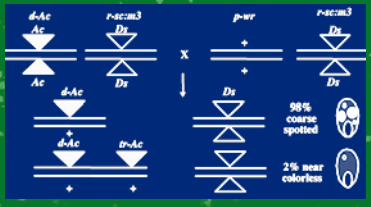
Methods in Molecular Biology™

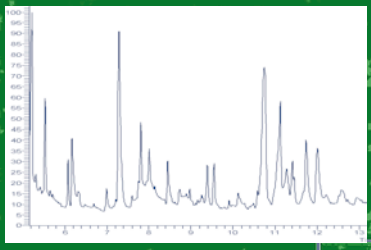
VOLUME 236

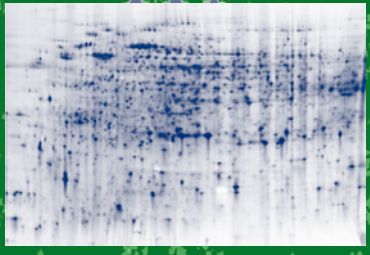
Plant Functional Genomics

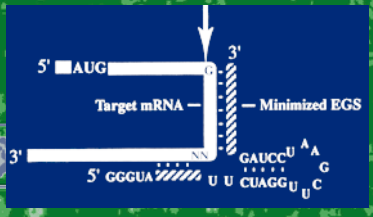
Edited by

Erich Grotewold









HUMANA PRESS

1

An Improved Method for Plant BAC Library Construction

Meizhong Luo and Rod A. Wing

Summary

Large genomic DNA insert-containing libraries are required as critical tools for physical mapping, positional cloning, and genome sequencing of complex genomes. The bacterial artificial chromosome (BAC) cloning system has become a dominant system over others to clone large genomic DNA inserts. As the costs of positional cloning, physical mapping, and genome sequencing continuously decrease, there is an increasing demand for high-quality deep-coverage large insert BAC libraries. In our laboratory, we have constructed many high-quality deep-coverage large insert BAC libraries including arabidopsis, manocot and dicot crop plants, and plant pathogens. Here, we present the protocol used in our laboratory to construct BAC libraries.

Key Words

BAC, library, method, pCUGIBAC1, plant

1. Introduction

Large genomic DNA insert-containing libraries are essential for physical mapping, positional cloning, and genome sequencing of complex genomes. There are two principal large insert cloning systems that are constructed as yeast or bacterial artificial chromosomes (YACs and BACs, respectively). The YAC cloning ***(1)*** was first developed in 1987 and uses *Saccharomyces cerevisiae* as the host and maintains large inserts (up to 1 Mb) as linear molecules with a pair of yeast telomeres and a centromere. Although used extensively in the late 1980s and early 1990s, this system has several disadvantages ***(2,3)***. The recombinant DNA in yeast can be unstable. DNA manipulation is difficult and inefficient. Most importantly, a high level of chimerism, the clon-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

ing of two or more unlinked DNA fragments in a single molecule, is inherent within the YAC cloning system. These disadvantages impede the utility of YAC libraries, and subsequently, this system has been gradually replaced by the BAC cloning system introduced in 1992 ***(4)***.

The BAC cloning uses a derivative of the *Escherichia coli* F-factor as vector and *E. coli* as the host, making library construction and subsequent downstream procedures efficient and easy to perform. Recombinant DNA inserts up to 200 kb can be efficiently cloned and stably maintained in *E. coli*. Although the insert size cloning capacity is much lower than that of the YAC system, it is this limited cloning capacity that helps to prevent chimerism, because the inserts with sizes between 130–200 kb can be selected, while larger inserts, composed of two or more DNA fragments, are beyond the cloning capacity of the BAC system or are much less efficiently cloned.

In 1994, our laboratory was the first to construct a BAC library for plants using *Sorghum bicolor* ***(5)***. Since then, we have constructed a substantial number of deep coverage BAC libraries, including *Arabidopsis* ***(6)***, rice ***(7)***, melon ***(8)***, tomato ***(9)***, soybean ***(10)***, and barley ***(11)*** and have provided them to the community for genomics research ([http://www.genome.arizona.edu] and [http://www.genome.clemson.edu]).

The construction of a BAC library is quite different from that of a general plasmid or λ DNA library used to isolate genes or promoter sequences by positive screening. Megabase high molecular weight DNA is required for BAC library construction. Because individual clones of the BAC library will be picked, stored, arrayed on filters, and directly used for mapping and sequencing, a BAC library with a small average insert size and high empty clone (no inserts) rate will dramatically increase the cost and labor for subsequent work. Usually, a BAC library with an average insert size smaller than 130 kb and empty clone rate higher than 5% is unacceptable. These strict requirements make BAC library construction much more difficult than the construction of a general DNA library.

As the costs of positional cloning, physical mapping, and genome sequencing continuously decrease, so increases the demand for high-quality deep-coverage large insert BAC libraries ***(12)***. As a consequence, we describe in this chapter how our laboratory constructs BAC libraries.

Several protocols have been published for the construction of high quality plant and animal BAC libraries ***(13–18)***, including three from our laboratory ***(16–18)***. We improved on these methods in several ways ***(8)***. First, to easily isolate large quantities of single copy BAC vector, pIndigoBAC536 (*see* **Note 1**) was cloned into a high copy cloning vector, pGEM-4Z. This new vector, designated pCUGIBAC1 (**Fig. 1**), replicates as a high copy vector and can be isolated in large quantity using standard plasmid DNA isolation methods. It

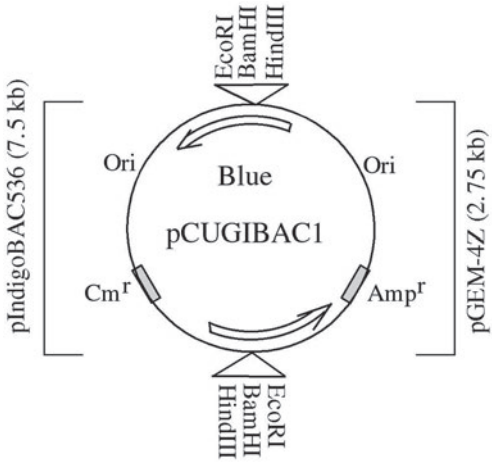


Fig. 1. pCUGIBAC1. Not drawn to scale.

retains all three unique cloning sites (*Hin*dIII, *Eco*RI, and *Bam*HI), as well as the two *Not*I sites flanking the cloning sites, of the original pIndigoBAC536. Second, to improve the stability of megabase DNA and size-selected DNA fractions in agarose, as well as digested dephosphorylated BAC vectors, we determined that such material can be stored indefinitely in 70% ethanol at –20°C and in 40–50% glycerol at –80°C, respectively.

The vector has been distributed to many users worldwide, and the high molecular weight DNA preservation method, established by Luo et al. ***(8)***, has been extensively used by colleagues and visitors and shown to be very efficient ***(18)***. These improvements and protocols described here save on resources, cost, and labor, and also release time constraints on BAC library construction.

2. Materials, Supplies, and Equipment

2.1. For pCUGIBAC1 Plasmid DNA Preparation

1. pCUGIBAC1 (http://www.genome.clemson.edu).
2. LB medium; 10 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl.
3. Ampicillin and chloramphenicol (Fisher Scientific).
4. Qiagen plasmid midi kit (Qiagen).
5. Thermostat shaker (Barnstead/Thermolyne).

2.2. For BAC Vector pIndigoBAC536 Preparation

2.2.1. For Method One

1. Restriction enzymes (New England Biolabs).
2. HK phosphatase, Tris-acetate (TA) buffer, 100 m*M* $CaCl_2$, ATP, T4 DNA ligase (Epicentre).

3. Agarose and glycerol (Fisher Scientific).
4. 10× Tris-borate EDTA (TBE) and 50× Tris-acetate EDTA (TAE) buffer (Fisher Scientific).
5. 1 kb DNA ladder (New England Biolabs).
6. Ethidium bromide (EtBr) (10 mg/mL).
7. λ DNA (Promega).
8. Water baths.
9. CHEF-DR III pulse field gel electrophoresis system (Bio-Rad).
10. Dialysis tubing (Spectra/Por2 tubing, 25 mm; Spectrum Laboratories).
11. Model 422 electro-eluter (Bio-Rad).
12. Minigel apparatus Horizon 58 (Whatman).
13. UV transilluminator.

2.2.2. For Method Two

1. Restriction enzymes and calf intestinal alkaline phosphatase (CIP) (New England Biolabs).
2. 0.5 *M* EDTA, pH 8.0.
3. Absolute ethanol, agarose, and glycerol (Fisher Scientific).
4. T4 DNA ligase (Promega).
5. 10× TBE and 50× TAE buffer (Fisher Scientific).
6. 1 kb DNA ladder.
7. EtBr (10 mg/mL).
8. λ DNA.
9. Water baths.
10. CHEF-DR III pulse field gel electrophoresis system.
11. Dialysis tubing (Spectra/Por2 tubing, 25 mm).
12. Model 422 electro-eluter.
13. Minigel apparatus Horizon 58.
14. UV transilluminator.

2.3. For Preparation of Megabase Genomic DNA Plugs from Plants

1. Nuclei isolation buffer (NIB): 10 m*M* Tris-HCl, pH 8.0, 10 m*M* EDTA, pH 8.0, 100 m*M* KCl, 0.5 *M* sucrose, 4 m*M* spermidine, 1 m*M* spermine.
2. NIBT: NIB with 10% Triton® X-100.
3. NIBM: NIB with 0.1% β-mercaptoethanol (add just before use).
4. Low melting temperature agarose (FMC).
5. Proteinase K solution: 0.5 *M* EDTA, 1% N-lauroylsarcosine, adjust pH to 9.2 with NaOH; add proteinase K to 1 mg/mL before use.
6. 50 m*M* phenylmethylsulfonyl fluoride (PMSF) (Sigma) stock solution (prepared in ethanol or isopropanol).
7. $T_{10}E_{10}$ (10 m*M* Tris-HCl and 10 m*M* EDTA, pH 8.0) and TE (10 m*M* Tris-HCl and 1 m*M* EDTA, pH 8.0).
8. Mortars, pestles, liquid nitrogen, 1-L flasks, cheese cloth, small paintbrush, and Pasteur pipet bulbs.

9. 50-mL Falcon® tubes (Fisher Scientific) and miracloth (Calbiochem-Novabiochem).
10. Plug molds (Bio-Rad).
11. GS-6R centrifuge (Beckman).
12. Model 230300 Bambino hybridization oven (Boekel Scientific).

2.4. For Preparation of High Molecular Weight Genomic DNA Fragments

2.4.1. For Pilot Partial Digestions

1. Restriction enzymes and BSA (Promega).
2. 40 m*M* Spermidine (Sigma) and 0.5 *M* EDTA, pH 8.0.
3. λ Ladder pulsed field gel (PFG) marker (New England Biolabs).
4. Agarose and 10× TBE.
5. EtBr (10 mg/mL).
6. Razor blades, microscope slides, and water baths.
7. CHEF-DR III pulse field gel electrophoresis system.
8. UV transilluminator.
9. EDAS 290 image system (Eastman Kodak).

2.4.2. For DNA Fragment Size Selection

1. Restriction enzymes and BSA.
2. 40 m*M* spermidine and 0.5 *M* EDTA, pH 8.0.
3. λ Ladder PFG marker.
4. Agarose and 10× TBE.
5. Low melting temperature agarose.
6. EtBr (10 mg/mL) and 70% ethanol.
7. Razor blades, microscope slides, water baths, and a ruler.
8. CHEF-DR III pulse field gel electrophoresis system.
9. UV transilluminator.
10. EDAS 290 image system.

2.5. For BAC Library Construction

2.5.1. For DNA Ligation

1. T4 DNA ligase and λ DNA.
2. Agarose and 1× TAE buffer.
3. EtBr (10 mg/mL).
4. Dialysis tubing (Spectra/Por2 tubing, 25 mm) or Model 422 electro-eluter.
5. Minigel apparatus Horizon 58.
6. UV transilluminator.
7. Water baths.
8. 0.1 *M* Glucose/1% agarose cones: melt 0.1 *M* glucose and 1% agarose in water, dispense 1 mL to each 1.5-mL microcentrifuge, insert a 0.5-mL microcentrifuge

into each 1.5-mL microcentrifuge containing 0.1 *M* glucose and 1% agarose, after solidification, pull out the 0.5-mL microcentrifuges.

2.5.2. For Test Transformation

1. DH10B T1 phage-resistant cells (Invitrogen).
2. SOC: 20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 m*M* NaCl, 2.5 m*M* KCl, autoclave, and add filter-sterilized $MgSO_4$ to 10 m*M*, $MgCl_2$ to 10 m*M*, and glucose to 20 m*M* before use.
3. 100-mm diameter Petri dish agar plates containing LB with 12.5 μg/mL of chloramphenicol, 80 μg/mL of x-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside or 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside [X-gal]) and 100 μg/mL of IPTG isopropyl-β-D-thiogalactoside or isopropyl-β-D thiogalactopyranoside.
4. 15-mL culture tubes.
5. Thermostat shaker.
6. Electroporator (cell porator; Life Technologies).
7. Electroporation cuvettes (Whatman).
8. 37°C incubator.

2.5.3. For Insert Size Estimation

2.5.3.1. For BAC DNA Isolation

1. LB with 12.5 μg/mL chloramphenicol.
2. Isopropanol and ethanol.
3. P_1, P_2, and P_3 buffers from plasmid kits (Qiagen).
4. 15-mL culture tubes.
5. Thermostat shaker.
6. Microcentrifuge.

2.5.3.2. For BAC Insert Size Analysis

1. *Not*I (New England Biolabs).
2. DNA loading buffer: 0.25% (w/v) bromophenol blue and 40% (w/v) sucrose in TE, pH 8.0.
3. MidRange I PFG molecular weight marker (New England Biolabs).
4. Agarose, 0.5× TBE buffer, and EtBr (10 mg/mL).
5. 37°C water bath or incubator.
6. CHEF-DR III pulse field gel electrophoresis system.
7. UV transilluminator.
8. EDAS 290 image system.

2.5.4. For Bulk Transformation, Colony Array, and Library Characterization

1. Freezing media: 10 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl, 36 m*M* K_2HPO_4, 13.2 m*M* KH_2PO_4, 1.7 m*M* Na-citrate, 6.8 m*M* $(NH_4)_2SO_4$,

4.4% glycerol, autoclave, and add filter-sterilized $MgSO_4$ stock solution to 0.4 m*M*.
2. 384-well plates and Q-trays (Genetix).
3. Toothpicks (hand picking) or Q-Bot (Genetix).

3. Methods

3.1. Preparing pCUGIBAC1 Plasmid DNA

1. Inoculate a single well-isolated *E. coli* clone harboring pCUGIBAC1 in LB containing 50 mg/L of ampicillin and 12.5 mg/L of chloramphenicol and grow at 37°C for about 20 h with continuous shaking.
2. Prepare pCUGIBAC1 plasmid DNA using the plasmid midi kit according to the manufacturer's instruction, except that after adding solution P_2, the sample was incubated at room temperature for not more than 3 min instead of 5 min (*see* acknowledgments). Each 100 mL of culture yields about 100 μg of plasmid DNA when using a midi column.

3.2. Preparing BAC Vector, pIndigoBAC536

3.2.1. Method One

1. Set up 4–6 restriction digestions, each digesting 5 μg pCUGIBAC1 plasmid DNA (with *Hin*dIII, *Eco*RI, or *Bam*HI depending on which enzyme is selected for BAC library construction) in 150 μL 1× TA buffer at 37°C for 2 h. Check 1 μL on a 1% agarose minigel to determine if the plasmid is digested.
2. Heat the digestions at 75°C for 15 min to inactivate the restriction enzyme.
3. Add 8 μL of 100 m*M* $CaCl_2$, 1.5 μL of 10× TA buffer, and 5 μL of HK phosphatase, and incubate the samples at 30°C for 2 h.
4. Heat the samples at 75°C for 30 min to inactivate the HK phosphatase.
5. Add 6.4 μL of 25 m*M* ATP, 5 μL of 2 U/μL T4 DNA ligase, and 1.3 μL of 10× TA buffer, incubate at 16°C overnight for self-ligation.
6. Heat the self-ligations at 75°C for 15 min.
7. Combine the samples and run the combined sample in a single well, made by taping together several teeth of the comb according to the sample vol, on a 1% CHEF agarose gel at 1–40 s linear ramp, 6 V/cm, 14°C in 0.5× TBE buffer along with the 1 kb ladder loaded into the wells on the both sides of the gel as marker for 16–18 h.
8. Stain the two sides of the gel containing the marker and a small part of the sample with 0.5 μg/mL EtBr and recover the gel fraction containing the 7.5-kb pIndigoBAC536 DNA band from the unstained center part of the gel by aligning it with the two stained sides. Undigested circular plasmid DNA and non-dephosphorylated linear DNA that has recircularized or formed concatemers after self-ligation should be reduced to an acceptable level after this step. **Figure 2** shows a gel restained with 0.5 μg/mL EtBr after having recovered the gel fraction containing the 7.5-kb pIndigoBAC536 vector. The 2.8-kb band is the pGEM-4Z vector.

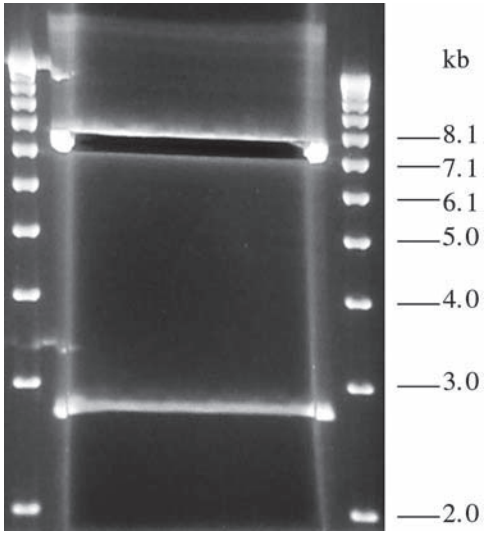


Fig. 2. Recovering linearized dephophorylated 7.5-kb pIndigoBAC536 vector from a CHEF agarose gel. See text for details.

9. Electroelute pIndigoBAC536 from the agarose gel slice in 1× TAE buffer at 4°C. Either dialysis tubing ***(19)*** or the Model 422 electro-eluter can be used ***(18)***.
10. Estimate the DNA concentration by running 2 μL of its dilution along with 2 μL of each of serial dilutions of λ DNA standards (1, 2, 4, and 8 ng/μL) on a 1% agarose minigel containing 0.5 μg/mL EtBr (for 10 min) and comparing the images under UV light, or simply by spotting a 1-μL dilution along with 1 μL of each of serial dilutions of λ DNA standards (1, 2, 4, and 8 ng/μL) on a 1% agarose plate containing 0.5 μg/mL EtBr and comparing the images under UV light after being incubated at room temperature for 10 min.
11. Adjust DNA concentration to 5 ng/μL with glycerol (final glycerol concentration 40–50%), aliquot into microcentrifuge tubes, and store the aliquots at –80°C. Use each aliquot only once.
12. Test the vector quality by cloning λ DNA fragments digested with the same restriction enzyme as used for vector preparation. Prepare a sample without the λ DNA fragments as the self-ligation control. For ligation, transformation, and insert check, follow the protocols in **Subheading 3.5**. for BAC library construction, except that inserts are checked on a standard agarose gel instead of a CHEF gel. Colonies from the ligation with the λ DNA fragments should be at least 100 times more abundant than those from the self-ligation control. More than 95% of the white colonies from the ligation with the λ DNA fragments should contain inserts.

3.2.2. Method Two

1. Set up 4–6 digestions, each digesting 5 μg pCUGIBAC1 plasmid DNA (with *Hin*dIII, *Eco*RI, or *Bam*HI depending on which enzyme is selected for BAC library construction) in 150 μL 1× restriction buffer at 37°C for 1 h. Check 1 μL on a 1% agarose minigel to see if the plasmid is digested.
2. Add 1 U of CIP and incubate the samples at 37°C for an additional 1 h (*see* **Note 2**).
3. Add EDTA to 5 m*M* and heat the samples at 75°C for 15 min.
4. Precipitate DNA with ethanol, wash it with 70% ethanol, air-dry, and add: 88 μL of water, 10 μL of 10× T4 DNA ligase buffer, and 2 μL of 3 U/μL T4 DNA ligase.
5. Incubate the samples at 16°C overnight for self-ligation. Then follow **steps 6–12** of Method One (**Subheading 3.2.1.**).

3.3. Preparing Megabase Genomic DNA Plugs from Plants (see [18] for alternatives) (see Note 3)

1. Young seedlings of monocotyledon plants, such as rice and maize, and young leaves of dicotyledon plants, such as melon, are used fresh or collected and stored at –80°C.
2. Grind about 100 g of tissue in liquid N_2 with a mortar and a pestle to a level that some small tissue chunks can be still seen (*see* **Note 4**).
3. Divide and transfer the ground tissue into two 1-L flasks, each containing 500 mL of ice-cold NIBM (1 g tissue/10 mL).
4. Keep the flasks on ice for 15 min with frequent and gentle shaking.
5. Filter the homogenate through four layers of cheese cloth and one layer of miracloth. Squeeze the pellet to allow maximum recovery of nuclei-containing solution.
6. Filter the nuclei-containing solution again through one layer of miracloth.
7. Add 1:20 (in vol) of NIBT to the nuclei-containing solution and keep the mixture on ice for 15 min with frequent and gentle shaking.
8. Transfer the mixture into 50-mL Falcon tubes. Centrifuge the tubes at 2400*g* at 4°C for 15 min.
9. Gently resuspend the pellets in the residual buffer by tapping the tubes or with a small paintbrush.
10. Dilute the nucleus suspension with NIBM and combine it into two 50-mL Falcon tubes. Adjust the vol to 50 mL with NIBM in each tube and centrifuge the tubes at 2400*g* at 4°C for 15 min.
11. Resuspend the pellets as in **step 9**. Dilute the nucleus suspension with NIBM and combine it into one 50-mL Falcon tube. Adjust the vol to 50 mL with NIBM and centrifuge it at 2400*g* at 4°C for 15 min.
12. Remove the supernatant and gently resuspend the pellet in approx 1.5 mL of NIB.
13. Incubate the nucleus suspension at 45°C for 5 min. Gently mix it with an equal vol of 1% low melting temperature agarose, prepared in NIB and pre-incubated

at 45°C, by slowly pipeting 2 or 3 times. Transfer the mixture to plug molds and let stand on ice for about 30 min to form plugs.
14. Tranfer <50 agarose plubs into each 50-mL Falcon tube, containing 40 mL of proteinase K solution, with a Pasteur pipet bulb.
15. Incubate the tubes in a hybridization oven (e.g., Model 230300 Bambino hybridization oven) at 50°C with a gentle rotation for about 24 h.
16. Repeat **step 15** with fresh proteinase K solution.
17. Wash the plugs, each time for about 1 h at room temperature with gentle shaking or rotation, twice with $T_{10}E_{10}$ containing 1 m*M* PMSF and twice with TE (40 mL each time for each 50-mL Falcon tube containing <50 plugs).
18. Store the plugs in TE buffer at 4°C (for frequent use) or rinse them with 70% ethanol and store in 70% ethanol (40 mL for each 50-mL Falcon tube containing <50 plugs) at –20°C (for long-term storage) (*see* **Note 5**).

3.4. Preparing High Molecular Weight Genomic DNA Fragments

3.4.1. Pilot Partial Digestions

1. Soak required number (e.g., 4 plugs) of TE-stored plugs in sterilized distilled water (more than 20 vol) for 1 h before partial digestion. For ethanol-stored plugs, transfer required number of 70% ethanol-stored plugs into TE buffer or directly into sterilized distilled water (more than 20 vol) at 4°C the day before use (*see* **Note 6**) and soak them in sterilized distilled water (more than 20 vol) for 1 h before partial digestion.
2. Dispense 45 µL of buffer mixture (24.5 µL of water, 9.5 µL of 10× restriction enzyme buffer, 1 µL of 10 mg/mL bovine serum albumin BSA, and 10 µL of 40 m*M* spermidine) into each of an ordered serial set (e.g., Nos. 1–8) of microcentrifuge tubes. Keep the microcentrifuge tubes on ice.
3. Chop each half DNA plug to fine pieces with a razor blade on a clean microscope slide (assume each half DNA plug has a vol of 50 µL) and transfer these pieces into a microcentrifuge tube containing 45 µL of restriction enzyme buffer on ice with a spatula. Mix by tapping and incubate on ice for 30 min.
4. Make serial dilutions of restriction enzyme (*Hin*dIII, *Eco*RI, or *Bam*HI, depending on which enzyme is selected for BAC library construction) with 1× restriction enzyme buffer (e.g., 0.4, 0.8, 1.2, 1.6, 2.0, and 2.4 U/µL).
5. Add 5 µL of one enzyme dilution to each of the microcentrifuge tube in **step 3**. Set up an uncut control, by not adding any enzyme, and a completely cut control, by adding 50–60 U of enzyme. Mix by tapping and incubate on ice for 30 min to allow for diffusion of the enzyme into the agarose matrix.
6. Incubate the microcentrifuge tubes in a 37°C water bath for 40 min.
7. Add 10 µL of 0.5 *M* EDTA, pH 8.0, to each microcentrifuge tube. Mix by tapping and incubate on ice for at least 10 min to terminate the digestions.
8. Prepare a 14 × 13 cm CHEF agarose gel by pouring 130 mL of 1% agarose (in 0.5× TBE buffer) at about 50°C into a 14 × 13 cm gel casting stand (Bio-Rad). Use two 15-well 1.5-mm-thick combs (Bio-Rad) bound together with tape for the samples. Set aside several milliliters of 1% agarose (in 0.5× TBE buffer) at 65°C.

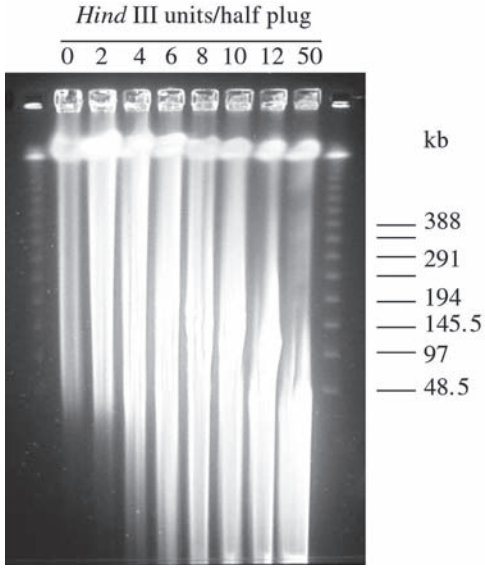


Fig. 3. Partial digestions of DNA plugs with serial dilutions of *Hin*dIII at 37°C for 40 min. DNA was separated on 1% CHEF agarose gel at 1–50 s linear ramp, 6 V/cm, 14°C in 0.5× TBE buffer for 20 h. The marker used is λ ladder PFG.

9. Load each sample from **step 7** into the center wells of the agarose gel with a spatula. Load the λ ladder PGF marker into the wells on the two sides of the gel. Seal the wells with the 1% agarose reserved at 65°C.
10. Run the gel at 1–50 s linear ramp, 6 V/cm, 14°C in 0.5× TBE buffer for 18–20 h.
11. Stain the gel with 0.5 μg/mL EtBr and take a photograph (*see* **Note 7**). **Figure 3** shows an example for the partial digestions of DNA plugs with serial dilutions of *Hin*dIII at 37°C for 40 min.

3.4.2. DNA Fragment Size Selection

1. Soak required number of plugs (e.g., 6 plugs) as in **Subheading 3.4.1.**, **step 1**.
2. Prepare a buffer mixture and dispense it into a set of microcentrifuge tubes (12 microcentrifuge tubes for 6 plugs) as in **Subheading 3.4.1.**, **step 2**.
3. Chop each half plug and treat the chopped plug pieces as in **Subheading 3.4.1.**, **step 3**.
4. Make the restriction enzyme dilution that produced the most DNA fragments in the range of 100–400 kb in the pilot partial digestion. For the batch of DNA plugs used in **Fig. 3**, 0.8 U/μL *Hin*dIII dilution (4 U of *Hin*dIII per half plug when 5 μL is used) was used for DNA fragment preparation.

5–7. Follow **Subheading 3.4.1.**, **steps 5–7**, except that 5 μL of the same enzyme dilution prepared in **step 4** is added to each of the microcentrifuge tubes in **step 3**.

8. Prepare a 14 × 13 cm CHEF agarose gel by pouring 130 mL of 1% agarose in

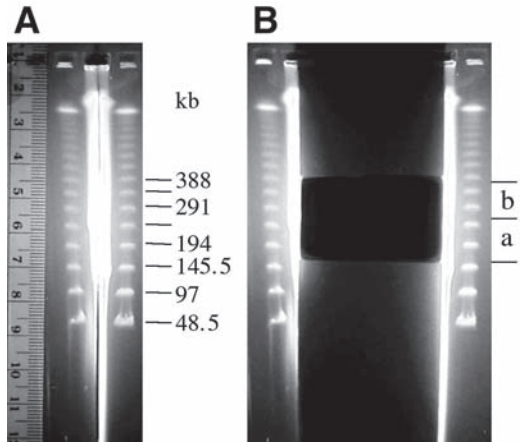


Fig. 4. An example for the first size selection of genomic DNA fragments. **(A)** Staining the two sides of the gel and taking a photograph with a ruler. **(B)** Recovering two gel fractions from the unstained center part of the gel corresponding to 150–250 and 250–350 kb located by a ruler.

0.5× TBE buffer at about 50°C into a 14 × 13 cm gel casting stand. Use a trimmed comb made by taping together several teeth of two 15-well 1.5-mm-thick combs to make a single well for the sample according to the sample vol.

9. Load the samples from **step 7** into the well with a spatula. Load the λ ladder PFG marker into the individual wells on the two sides of the gel. Seal the wells with 1% agarose in 0.5× TBE buffer maintained at 65°C.
10. Run the gel at 1–50 s linear ramp, 6 V/cm, 14°C in 0.5× TBE buffer for 18–20 h.
11. Stain the two sides of the gel containing the marker and a small part of the sample with 0.5 μg/mL EtBr and take a photograph with a ruler at one side (**Fig. 4A**).
12. Recover two gel fractions (first size-selected fractions: a and b) from the unstained center part of the gel corresponding to 150–250 and 250–350 kb located by a ruler (**Fig. 4B**).
13. Place the two gel fractions side by side (with a gap between them) on the top of a 14 × 13 cm gel casting stand with the orientation the same as in the original gel in **step 12**. Pour 130 mL of 1% agarose in 0.5× TBE at about 50°C into the gel casting stand to form a second gel encasing the two gel factions.
14. Run the gel at 4 s constant time, 6 V/cm, 14°C in 0.5× TBE buffer for 18–20 h.
15. Stain the two sides with 0.5 μg/mL EtBr, each containing a small part of one of the two first size-selected fractions, and the center part that contains the small parts of both first size-selected fractions. Take a photograph with a ruler at one side.
16. For each first size-selected fraction (a and b), recover two gel fractions (second size-selected fractions: a1 and a2, and b1 and b2) located by a ruler (**Fig. 5**). Gel fractions are used immediately or stored at –20°C in 70% ethanol (*see* **Note 5**).

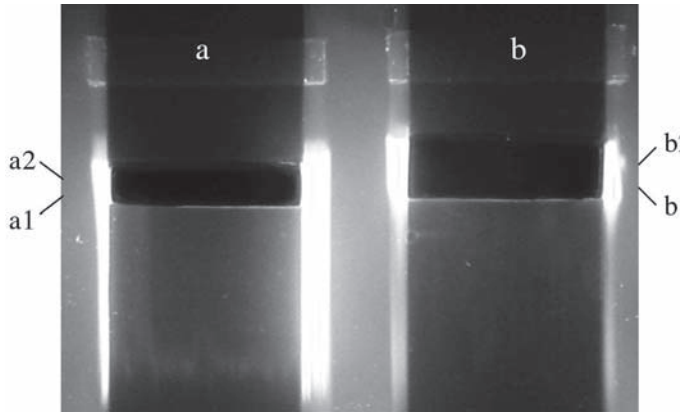


Fig. 5. An example for the second size selection of genomic DNA fragments.

3.5. BAC Library Construction

3.5.1. DNA Ligation

1. Transfer required amount of each 70% ethanol-stored fraction (e.g., one-third to one-half fraction) into 1× TAE buffer (more than 20 vol) at 4°C the day before use (*see* **Note 8**).
2. Electroelute high molecular weight genomic DNA at 4°C from fresh gel fractions or 1× TAE buffer soaked 70% ethanol-stored fractions in 1× TAE buffer. Either dialysis tubing ***(20)*** or Model 422 electro-eluter ***(18)*** can be used. Eluted DNA should be used as soon as possible (use it the same day it is eluted). Always use pipet tips with the tips cut off when manipulating high molecular weight genomic DNA to avoid mechanical shearing.
3. Estimate the DNA concentration by running 5 μL of the eluted DNA along with 2 μL of serial dilutions of λ DNA standards (1, 2, 4, 8, and 16 ng/μL) on a 1% agarose minigel containing 0.5 μg/mL EtBr (for 10 min) and comparing the images under UV light.
4. Set up ligations: in each microcentrifuge tube, add 4 μL of 5 ng/μL vector and 84 μL of DNA eluted in 1× TAE containing up to 200 ng of high molecular weight genomic DNA fragments. If the eluted DNA has a high concentration, dilute it with sterilized water. Incubate the vector–genomic DNA fragment mixture at 65°C for 15 min, cool at room temperature for about 10 min, and add 10 μL of 10× T_4 DNA ligase buffer and 2 μL of 3 U/μL T_4 DNA ligase. Incubate the ligations at 16°C overnight.
5. Heat the ligations at 65°C for 15 min to terminate the ligation reactions.
6. Transfer ligation samples into 0.1 *M* glucose/1% agarose cones (*see* **Subheading 2.5.1.**) to desalt for 1.5 h on ice ***(20)*** or transfer ligation samples onto filters (Millipore) floating on 5% polyethylene glycol (PEG)8000 in Petri dishes set on ice for 1.5 h as modified from Osoegawa et al. ***(15)***. Store the ligations at 4°C for not more than 10 d.

3.5.2. Test Transformation

1. Thaw ElectroMax DH10B T_1 phage-resistant competent cells on ice and dispense 16 μL into prechilled microcentrifuge tubes on ice. Precool the electroporation cuvettes on ice. Prepare SOC media and dispense 0.5 mL to each sterile 15-mL culture tube. Label the microcentrifuge tubes, cuvettes, and culture tubes coordinately.
2. Take 1 to 2 μL of ligated DNA from each ligation sample and mix it with the competent cells by gentle tapping.
3. Transfer the DNA/competent cell mixture from each microcentrifuge tube into precooled electroporation cuvettes. Electroporate on ice at 325 DC V with fast charge rate at a low resistance (4 kΩ) and a capacitance of 330 μF. We did not find a significant difference when different DC V between 300–350 V were applied.
4. Transfer the electroporated cells from each cuvette into sterile 15-mL culture tubes containing 0.5 mL SOC. Incubate the cultures at 37°C for 1 h with vigorous shaking.
5. Plate 20 and 200 μL of each culture on 100-mm diameter Petri dish agar plates containing LB with 12.5 μg/mL of chloramphenicol, 80 μg/mL X-gal, and 100 μg/mL IPTG. Incubate the plates at 37°C overnight.
6. Count the white colonies and determine the number of recombinant clones per microliter of ligation. This number, the genome size, and the required genome coverage will be considered to decide if the experiment should be continued. For example, 3 parallel 100 μL ligations of 100 white colonies/μL with the expected average insert size of 130 kb will result in about 9 genome coverages for rice (genome size is 430 Mbp), but only 1.56 genome coverages for maize (genome size is 2500 Mbp).

3.5.3. Insert Size Estimation

3.5.3.1. BAC DNA Isolation

Several automated methods, such as using an Autogen 740 (AutoGen) or using a Quadra 96 (TomTec) can be used to isolate BAC DNA. A manuscript for a detailed method for preparing BAC DNA with a Quadra 96 is in preparation by HyeRan Kim et al. Here we present a manual method adapted from the Qiagen method.

1. Randomly pick white colonies with sterilized toothpicks and inoculate each into 2 mL of LB containing 12.5 μg/mL chloramphenicol in a sterile 15-mL culture tube. Grow the cells at 37°C overnight with vigorous shaking.
2. Transfer each cell culture (about 1.5 mL) into a microcentrifuge tube and collect cells at 16,000*g* (at room temperature or 4°C) for 10 min; remove supernatant.
3. Add 200 μL of P_1. Mix the tubes with a vortex to resuspend pellets at room temperature.

4. Add 200 μL of P_2. Mix the contents gently but thoroughly by inverting the tubes 3 to 4 times. Stand the tubes at room temperature for not more than 3 min.
5. Add 200 μL of P_3. Mix the contents gently but thoroughly by inverting the tubes 3 to 4 times. Stand the tubes on ice for 15 min.
6. Centrifuge the samples at 16,000*g* (at room temperature or 4°C) for 30–40 min.
7. Carefully transfer about 550 μL of each supernatant to a new microcentrifuge tube containing 400 μL of isopropanol. Mix the contents gently.
8. Centrifuge the samples at 16,000*g* (at room temperature or 4°C) for 30 min.
9. Remove the supernatant. Add 400 μL of 70% ethanol and centrifuge the samples at 16,000*g* for 10 min to wash the DNA pellets.
10. Remove the supernatant carefully with a pipet. Air-dry the DNA pellets, and resuspend in 60 μL of TE buffer, pH 8.0.

3.5.3.2. BAC Insert Size Analysis

1. Dispense 11 μL of *Not*I digestion mixture (8.85 μL of water, 1.5 μL of 10× buffer, 0.15 μL of 10 mg/mL BSA, and 0.5 μL of 10 U/μL *Not*I) into each microcentrifuge tube or each well of a 96-well microtiter plate.
2. Add 4 μL of BAC plasmid DNA to each tube or each well. Spin the samples briefly. Incubate the samples at 37°C for 3 h. Dispense 3 μL of 6× DNA loading buffer ***(21)*** into each tube or each well. Spin the samples briefly.
3. Prepare a 21 × 14 cm CHEF agarose gel by pouring 150 mL of 1% agarose in 0.5× TBE buffer at about 50°C into a 21 × 14 cm gel casting stand. Use a 45-well 1.5-mm-thick comb for the samples.
4. Load DNA samples. Use MidRange I as the size marker.
5. Run the gel at 5–15 s linear ramp, 6 V/cm, 14°C in 0.5× TBE buffer for 16 h.
6. Stain the gel with 0.5 μg/mL EtBr. Take a photograph of the gel. Analyze the insert sizes.

3.5.4. Bulk Transformation, Colony Array, and Library Characterization

If the test colonies meet the requirement for average insert size and empty vector rate, transform all ligated DNA into ElectroMax DH10B T_1 phage-resistant competent cells. Pick individual colonies into wells of 384-well plates containing freezing media manually or robotically (Q-Bot) and characterize the BAC library by insert size analysis of random clones. Store the BAC library at –80°C.

4. Notes

1. pIndigoBAC536 has the same sequence as pBeloBAC11, except that the internal *Eco*R1 site was destroyed so that the unique *Eco*R1 site in the multiple cloning site can be used for cloning, and a random point mutation was selected for in the lac Z gene that provides darker blue colony color on X-gal/IPTG selection. The GenBank® accession number for pBeloBAC11 is U51113.
2. CIP is active in many different buffers.

3. Plug preparation is a critical part of the work for plant BAC library construction. Many failures are attributed to the plugs not containing enough megabase DNA. To increase the DNA content in plugs, more starting material can be used, and the resultant nuclei can be imbedded in fewer plugs. However, at least 25–35 plugs for each preparation are required for convenient subsequent manipulation. The same batch of plugs should be used for pilot partial digestion and scaled partial digestion for BAC library construction.
4. Do not grind the material to a complete powder, as novices in this field usually do. Overgrinding reduces the yield of nuclei dramatically.
5. Allow to stand at room temperature for about 30 min or at 4°C overnight before transferring to –20°C to avoid freezing the center part of the gel slices. Freezing causes high molecular weight DNA to shear.
6. If the 70% ethanol-stored plugs are needed to be used the same day, soak them in a large vol of sterilized distilled water (40 mL in a 50-mL Falcon tube) at room temperature for 3 h with gentle shaking and several changes of sterilized distilled water.
7. If the DNA in the completely cut control is not well digested (most of the DNA fragments should be below 50 kb after complete digestion), rewash the DNA plugs or use a different restriction enzyme. If a restriction condition to produce most of the DNA fragments in the range of 100–400 kb is not found, because of insufficient digestion or over digestion, repeat the pilot partial digestion with higher or lower enzyme concentrations respectively.
8. Similar to **Note 6**, if the 70% ethanol-stored fractions are needed to be used the same day, soak them in a large vol of 1× TAE buffer (40 mL in a 50-mL Falcon tube) at room temperature for 3 h with gentle shaking and several changes of 1× TAE buffer.

Acknowledgments

Jose Luis Goicoechea for BAC plasmid DNA preparation. We thank Dave Kudrna for his critical reading and suggestions.

References

1. Burke, D. T., Carle, G. F., and Olson, M. V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236,** 806–812.
2. Anderson, C. (1993) Genome shortcut leads to problems. *Science* **259,** 1684–1687.
3. Zhang, H. B. and Wing, R. A. (1997) Physical mapping of the rice genome with BACs. *Plant Mol. Biol.* **35,** 115–127.
4. Shizuya, H., Birren, B., Kim, U.-J., et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* **89,** 8794–8797.
5. Woo, S. S., Jiang, J., Gill, B. S., Paterson, A. H., and Wing, R. A. (1994) Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. *Nucleic Acids Res.* **22,** 4922–4931.

6. Choi, S. D., Creelman, R., Mullet, J., and Wing, R. A. (1995) Construction and characterization of a bacterial artificial chromosome library from *Arabidopsis thaliana*. *Weeds World* **2,** 17–20.
7. Chen, M., Presting, G., Barbazuk, W. B., et al. (2002) An integrated physical and genetic map of the rice genome. *Plant Cell* **14,** 537–545.
8. Luo, M., Wang, Y.-H., Frisch, D., Joobeur, T., Wing, R. A., and Dean, R. A. (2001) Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (*Fom-2*). *Genome* **44,** 154–162.
9. Budiman, M. A., Mao, L., Wood, T. C., and Wing, R. A. (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res.* **10,** 129–136.
10. Tomkins, J. P., Mahalingam, R., Smith, H., Goicoechea, J. L., Knap, H. T., and Wing, R. A. (1999) A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. *Plant Mol. Biol.* **41,** 25–32.
11. Yu, Y., Tomkins, J. P., Waugh, R., et al. (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare L.*) and the identification of clones containing putative resistance genes. *TAG* **101,** 1093–1099.
12. Couzin, J. (2002) NSF's ark draws alligators, algae, and wasps. *Science* **297,** 1638–1639.
13. Amemiya, C. T., Ota, T., and Litman, G. W. (1996) *Nonmammalian Genomic Analysis: A Practical Guide* (Lai, E. and Birren, B., eds.), Academic Press, San Diego, pp. 223–256.
14. Birren, B., Green, E. D., Klapholz, S., Myers, R. M., and Roskams, J. (eds.) (1997) *Analyzing DNA*. CSH Laboratory Press, Cold Spring Harbor, NY.
15. Osoegawa, K., Woon, P. Y., Zhao, B., et al. (1998) An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52,** 1–8.
16. Zhang, H. B., Woo, S. S., and Wing, R. A. (1996) *Plant Gene Isolation* (Foster, G. and Twell, D., eds.), John Wiley & Sons, New York, pp. 75–99.
17. Choi, S. and Wing, R. A. (2000) *Plant Molecular Biology Manual, 2nd ed.* (Gelvin, S. and Schilperoort, R., eds.), Kluwer Academic Publishers, Norwell, MA, pp. 1–28.
18. Peterson, D. G., Tomkins, J. P., Frisch, D. A., Wing, R. A., and Paterson, A. H. (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J. Agric. Genomics* **5,** (http://www.ncgr.org/jag).
19. Strong, S. J., Ohta, Y., Litman, G. W., and Amemiya, C. T. (1997) Marked improvement of PAC and BAC cloning is achieved using electroelution of pulsed-field gel-separated partial digests of genomic DNA. *Nucleic Acids Res.* **25,** 3959–3961.
20. Atrazhev, A. M. and Elliott, J. F. (1996) Simplified desalting of ligation reactions immediately prior to electroporation into *E. coli*. *BioTechniques* **21,** 1024.
21. Sambrook, J. and Russell, D. W. (eds.) (2001) *Molecular Cloning: A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.

2

Constructing Gene-Enriched Plant Genomic Libraries Using Methylation Filtration Technology

Pablo D. Rabinowicz

Summary

Full genome sequencing in higher plants is a very difficult task, because their genomes are often very large and repetitive. For this reason, gene targeted partial genomic sequencing becomes a realistic option. The method reported here is a simple approach to generate gene-enriched plant genomic libraries called methylation filtration. This technique takes advantage of the fact that repetitive DNA is heavily methylated and genes are hypomethylated. Then, by simply using an *Escherichia coli* host strain harboring a wild-type modified cytosine restriction (McrBC) system, which cuts DNA containing methylcytosine, repetitive DNA is eliminated from these genomic libraries, while low copy DNA (i.e., genes) is recovered. To prevent cloning significant proportions of organelle DNA, a crude nuclear preparation must be performed prior to purifying genomic DNA. Adaptor-mediated cloning and DNA size fractionation are necessary for optimal results.

Key Words

gene-enriched libraries, shotgun sequencing, Mcr, DNA methylation, retrotransposons, gene discovery, repetitive DNA

1. Introduction

Highly accurate full genomic sequencing like that performed for example in *Saccharomyces cerevisiae* ***(1)*** and *Caenorhabditis elegans* ***(2)*** has proven to be an invaluable resource to accelerate all areas of biological research. In particular in plants, the *Arabidopsis thaliana* genome sequence has been deciphered, meeting the highest standards of accuracy ***(3)***. Undoubtedly, the availability of this information had an immense impact not only in the *Arabidopsis* community, but in research in all other plant systems as well. Unfortunately, the production of such a high quality genomic resource is not an easy task. It implies

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

a significant amount of sequence redundancy only achievable by producing a huge number of sequence reads. Such reads are assembled and processed to produce as long contiguous stretches as possible, called contigs. In order to link these contigs in the right order and orientation, a large insert genomic library (using bacterial artificial chromosome [BAC] or P1-derived artificial chromosome [PAC] vectors) needs to be constructed, at least partially sequenced, and physically mapped.

A major obstacle to obtain the complete and accurate sequence of a complex (i.e., eukaryote) genome is the presence of large amounts of repetitive DNA. This DNA is composed of satellite DNA, transposons and retrotransposons, among other repeats, which often show a high degree of sequence conservation. For this reason, the computer software designed to assemble random sequence reads fails to build correct contigs of repetitive sequences, usually assembling most members of a repeat family in a single contig, regardless of their actual location in the genome.

In the early 1980s by the time the idea of sequencing the human genome was opened to discussion for the first time ***(4)***, Putney et al. ***(5)*** reported a method that allowed to discover new genes simply by cDNA sequencing, later called expressed sequence tag (EST) sequencing ***(6)***. This widely used technique allows obtaining gene sequence information getting around the problem of sequencing repetitive DNA. However, the EST approach has two main limitations. The first is the redundancy of cDNA libraries. Some cDNAs are often overrepresented and will be sequenced many times before a cDNA corresponding to a weakly expressed gene is found. The second limitation is the partial representation due to the tissue-specific and developmental regulation of gene expression. Some genes are expressed only in certain tissues or cells, and some are developmentally regulated. In order to recover the corresponding ESTs, libraries from several different tissues and developmental stages need to be constructed. Another although minor, disadvantage of EST sequencing is that repetitive elements are often transcribed and thus included in EST collections.

One way to solve the problem of the redundancy is to use normalized libraries ***(7)***. Normalization techniques are based on reassociation kinetics and have been improved to avoid the elimination of members of gene families. However, it is not trivial to obtain a normalized library where representation is acceptable. Regardless of these limitations, EST projects are being conducted for many organisms and are a key tool for gene discovery, annotation of genes, cross-species comparative analysis, and definition of intron–exon boundaries among many other uses. In particular for plants, ESTs have been the alternative to full genome sequence, because the genomes of many plants, often important crop species, are very large and repetitive. Usually, the genome size (or subgenome size in the case of polyploids) correlates with the proportion of

repetitive DNA. It has been proposed that all diploid higher plant genomes share essentially the same set of genes, called the "gene space" ***(8)***. Then, the bigger the genome, the higher sequencing cost per gene, due to the amount of nongenic (e.g., repetitive) DNA that needs to be sequenced before reaching a gene.

The conservation of coding sequences across different species allows identifying genes simply by comparing two different genomes. Frequently, gene modeling software fails to identify genes that can be spotted with this comparative genomics approach. Furthermore, once the complete genomic sequence is obtained for one organism, it can be compared to a draft (lowly redundant and discontinuous) sequence of a related organism. This approach yields a lot of new information for both species under analysis. The additional advantage of genomic vs cDNA sequencing in terms of representation makes the lowly redundant genomic sequencing a cost-effective process. In the case of plants however, the large genome sizes prevent the pursuit of full or even draft genomic sequencing projects. For these reasons, alternatives to obtain genomic sequences enriched in genes avoiding the repetitive DNA have been developed. In maize for example, the very active transposon *Mutator* ***(9)*** shows a strong bias to insert in low copy DNA (i.e., genes). By generating large *Mutator*-induced insertional mutagenesis, it is possible to collect genomic sequences flanking transposon insertion sites, which will mainly correspond to genes ***(10)***. Although *Mutator* insertions may not be completely at random in the genome, it can be a good complement to an EST project.

Another alternative for gene enriched genomic sequencing of plants is the methylation filtration technique, which takes advantage of the fact that most of the repetitive elements in plants are heavily methylated, while genes are hypomethylated. Because of their methylation status, repeats are sensitive to bacterial restriction-modification systems, in particular the Mcr system ***(11,12)***, which includes two restriction enzymes: McrA and McrBC. McrBC recognizes DNA containing 5-methylcytosine preceded by a purine ***(13)***. Restriction requires two of these sites separated by 40–2000 nucleotides. Such recognition sites are very frequent in any methylated genomic DNA. Thus, by the selecting a *mcrBC*$^{+}$ *Escherichia coli* host strain, repetitive DNA can be largely excluded from genomic shotgun libraries, preserving the low copy DNA. Basically, methylation filtration consists in shearing and size fractionation of genomic DNA to select fragments smaller than the estimated size of the genes. Larger fragments have a high probability of including some portion of repetitive DNA, which would be methylated and thus counter-selected in the filtered library. On the other hand, if fragments are too small, there are more chances to recover small fragments of repetitive DNA with low GC content. Such fragments may be poor in methylated sites susceptible to restriction by McrBC and

then can be frequently recovered in filtered libraries. The selected fragments are then end-repaired and cloned into a standard sequencing vector. Subsequently, the ligation is introduced in a *mcrBC*$^{+}$ *E. coli* host. The recombinant clones isolated after plating are picked for automatic sequencing. The same ligation mixture can be transformed into a *mcrBC*$^{-}$ *E. coli* strain to obtain an unfiltered control library.

The technique works very well for maize ***(14)***, and there is evidence that it works for many other plants (Rabinowicz and Martienssen, unpublished). The advantage of methylation-filtered libraries vs cDNA and transposon insertion libraries is that there is no bias towards a certain region of the genome or a given fraction of the genes. It is possible though, that methylated genes are not recovered in filtered libraries. However, gene methylation is often restricted to defined regions of the gene, mainly the ends ***(15–17)***. This would allow to clone at least most of the coding sequence of methylated genes. Furthermore, genes that are regulated by methylation may become demethylated during different developmental stages. In these cases, the construction of methylation-filtered libraries from a couple of developmental stages of a given plant would likely overcome the problem. For larger scale projects, another problem is posed by the cloning efficiency. In plants with very large genomes, repetitive DNA may account for more than 90% of the nuclear DNA. Then, most of the DNA is likely to be methylated leaving a very small fraction of the genome to be recovered in methylation-filtered libraries. As a result, the number of recombinant clones recovered after plating a filtered library may be <10% of the number of clones obtained in the corresponding unfiltered control library. Furthermore, the proportion of nonrecombinant background (blue colonies) may become significant. The use of adaptors often improves the cloning efficiency in addition to reduce the formation of chimerical clones. The cloning protocol presented here uses three-nucleotide overhang adaptors and a compatible sticky-end vector made by filling in one nucleotide in the four-nucleotide 5' overhang generated by a restriction nuclease ***(18)***. The advantage of using three- vs four-nucleotide overhang is that the nonrecombinant background is highly reduced because the vector ends become incompatible.

2. Materials

2.1. Nuclear DNA Preparation

1. Isolation buffer 1 (IB 1): 25 m*M* citric acid (pH to 6.5 with 1 *M* NaOH), 250 m*M* sucrose, 0.7% Triton® X-100, 0.1% 2-mercaptoethanol (*see* **Note 1**). IB 1 can be prepared at a 5× concentration. 2-Mercaptoethanol should be added immediately before usage.
2. Centrifuge tubes.
3. Liquid N_2.

4. Blender.
5. Polytron (Brinkmann Instruments).
6. Two 15-cm wide funnels.
7. Ring stand and clamps.
8. Cheesecloth (Fisher Scientific).
9. 60-μm Nylon mesh (Millipore).
10. 500-mL Centrifuge bottles with rubber o-ring sealing cap (Nalgene).
11. Isolation buffer 2 (IB 2): 50 m*M* Tris-HCl, pH 8.0, 25 m*M* EDTA, 350 m*M* sorbitol 0.1% 2-mercaptoethanol.
12. 5% Sarkosyl.
13. 5 *M* NaCl.
14. CTAB solution: 8.6% CTAB (Sigma), 0.7 *M* NaCl.
15. Chloroform:octanol (24:1).
16. Isopropanol.
17. 70% ethanol.
18. 10 m*M* Tris-HCl, pH 8.0.
19. Glass rod with bent tip.

2.2. DNA Shearing and End-Repairing

1. Glycerol 50%.
2. 10× Nebulization buffer: 0.5 *M* Tris-HCl, pH 8.0, 150 m*M* $MgCl_2$.
3. 14-mL Falcon® tubes (Becton Dickinson, cat. no. 35–2059).
4. Aero-mist nebulizer (CIS-US; cat. no. CA-209).
5. N_2 gas cylinder with a regulator able to deliver 1–50 psi.
6. Three-sixteenths-inch internal diameter PVC tubing (Fisher Scientific).
7. Parafilm.
8. 5 *M* NaCl.
9. Ethanol.
10. 70% Ethanol.
11. SpeedVac® (Savant Instruments).
12. 5 m*M* Tris-HCl, pH 8.0.
13. dNTPs 0.5 m*M* each (Roche Molecular Biochemicals).
14. T4 DNA polymerase (New England Biolabs).
15. T4 DNA polymerase buffer (New England Biolabs).
16. Klenow enzyme (Roche Molecular Biochemicals).
17. QIAquick™ polymerase chain reaction (PCR) purification kit (Qiagen).
18. T4 Polynucleotide kinase (PNK) (New England Biolabs).
19. T4 PNK buffer (New England Biolabs).
20. 100 m*M* ATP (Roche Molecular Biochemicals).
21. Equilibrated phenol:chloroform (1:1).

2.3. Adaptor Ligation

1. 200 μ*M* Top adaptor oligonucleotide 5'[P]-TAGACGCCTCGAG.
2. 200 μ*M* Bottom adaptor oligonucleotide 5'[OH]-CTCGAGGCGT.
3. 1 *M* NaCl.
4. T4 DNA ligase (Roche Molecular Biochemicals).
5. T4 DNA ligase buffer (Roche Molecular Biochemicals).
6. TEN buffer: 10 m*M* Tris-HCl, pH 7.5, 0.1 m*M* EDTA, 25 m*M* NaCl.
7. cDNA size fractionation columns (Invitrogen, Carlsbad, CA, USA).

2.4. Vector Preparation

1. Supercoiled pUC 19 DNA.
2. *Xba*I (Roche Molecular Biochemicals).
3. H buffer (Roche Molecular Biochemicals).
4. L buffer (Roche Molecular Biochemicals).
5. 10 mg/mL bovine serum albumin (BSA) (New England Biolabs).
6. 1 m*M* dCTP (Roche Molecular Biochemicals).
7. Klenow enzyme (Roche Molecular Biochemicals).
8. Calf intestinal phosphatase (CIP) (Roche Molecular Biochemicals).
9. CIP buffer (Roche Molecular Biochemicals).
10. 0.5 *M* EDTA.
11. Equilibrated phenol:chloroform (1:1).
12. QIAquick PCR purification kit.
13. Chloroform.
14. 5 *M* NaCl .
15. Ethanol.
16. 70% Ethanol.
17. 10 m*M* Tris-HCl, pH 8.0.

2.5. Preparation of Electrocompetent Cells

1. SOB medium without magnesium: 20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 2.5 m*M* KCl, and 0.5 g/L NaCl (pH 7.0 with NaOH, autoclaved).
2. 10% Glycerol (autoclaved).
3. Sterile 250-mL centrifuge bottles with rubber o-ring sealing cap.
4. Sterile 14-mL centrifuge tubes.

2.6. Electroporation

1. Electroporation cuvettes 0.1 cm (Bio-Rad).
2. Electroporator (Bio-Rad).
3. SOC medium: 20 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 2.5 m*M* KCl, and 0.5 g/L NaCl (pH 7.0 with NaOH, autoclaved, sterile 2 *M* $MgCl_2$, and 1 *M* glucose are added to a final concentration of 10 and 20 m*M*, respectively, after cooling down).
4. Sterile 14-mL centrifuge tubes.

5. Isopropyl β-D-thiogalactopyranoside (IPTG) 200 mg/mL.
6. 5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) 20 mg/mL in dimethylformamide.
7. LB-ampicillin agar plates: 10 g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl (pH 7.0 with NaOH); agar is added to a final concentration of 1.5%, autoclaved, cooled to 55°C, ampicillin is added to a final concentration of 100 μg/mL, and plates are poured).

2.7. Ligation

1. Ligation buffer.
2. Ligase (Roche Molecular Biochemicals).
3. 10 m*M* NaCl.
4. QIAquick PCR purification kit.

2.8. Checking the Average Library Insert Size by Colony PCR

1. 10× PCR buffer (Qiagen).
2. dNTP mixture (10 m*M* each dNTP) (Qiagen).
3. *Taq* DNA polymerase 5 U/μL (Qiagen).
4. 10 μ*M* M13/pUC sequencing (–40) primer (New England Biolabs).
5. 10 μ*M* M13/pUC reverse sequencing (–24) primer (New England Biolabs).
6. 250 μL PCR tubes or 8-strips (MJ Research).

3. Methods

3.1. Nuclear DNA Preparation

Plastids are very abundant, not only in green tissues, and their DNA is unmethylated. Thus, if chloroplast DNA is present in a DNA sample, it will be selected during the filtering process. For this reason, it is important to purify nuclei from the rest of the cell organelles before purifying the genomic DNA. The protocol used here is a modification of those reported by Kiss et al. and Wagner et al. ***(19,20)***.

1. In a cold room, prepare a ring stand with two funnels attached with clamps, one on top of the other, so that the top funnel drains inside the bottom one. Cover the upper funnel with four 30 × 30 cm layers of cheese cloth and the lower one with one 30 × 30 cm layer of 60-μm nylon mesh. Put a 500-mL centrifuge bottle under the lower funnel to collect the liquid.
2. Grind 50–100 g of frozen tissue in liquid N_2 (*see* **Note 2**).
3. Transfer to a blender containing 6–8 vol of IB 1.
4. Homogenize 3× at maximum speed for 10 s each time.
5. Transfer to a plastic beaker and further homogenize 3× with a polytron, 5 s each time (*see* **Note 3**).
6. Slowly pour the slurry into the top funnel.
7. When it stops dripping, squeeze the liquid out of the cheese cloth using gloves.

8. Centrifuge at 2000*g* for 15 min at 4°C.
9. Carefully discard the supernatant and resuspend the nuclear pellet in 0.1–0.5 vol of IB 1.
10. Transfer to 14- or 50-mL centrifuge tubes and centrifuge at 2000*g* for 15 min at 4°C.
11. Resuspend in 5–20 mL of IB 2.
12. Add one-fifth vol of 5% Sarkosyl.
13. Mix gently and incubate 15 min at room temperature.
14. Add one-seventh vol of 5 *M* NaCl and mix gently.
15. Add one-tenth vol of CTAB solution preheated to 60°C.
16. Mix gently and incubate for 30 min at 60°C, mixing by inversion every 2–4 min.
17. Add 1 vol of chloroform:octanol and mix well by inversion (do not vortex mix).
18. Centrifuge at 6000*g* for 15 min at 4°C.
19. Transfer upper phase to a new centrifuge tube.
20. Add two-thirds vol of isopropanol and mix slowly by inversion.
21. Hook the DNA with a glass rod bent in the tip to help preventing the DNA from falling off (*see* **Note 4**).
22. Wash the nuclear DNA by immersing the glass rod in 70% ethanol.
23. Air-dry the DNA for a few minutes.
24. Immerse the DNA in 0.5–1 mL 10 m*M* Tris-HCl, pH 8.0, and shake it quickly until it falls off the glass rod.
25. Let the DNA resuspend overnight at 4°C.

3.2. DNA Shearing and End-Repairing

1. In a 14-mL Falcon centrifuge tube, mix 20 μg of nuclear DNA with 1 mL of 50% glycerol and 0.2 mL of nebulization buffer. Add water up to a final vol of 2 mL.
2. Seal the bottom nebulizer inlet with parafilm.
3. Remove the nebulizer screw-cap and transfer the DNA mixture to the bottom of the nebulizer.
4. Put the nebulizer cap and attach N_2 gas tubing in the bottom inlet. Close the upper nebulizer outlet with the Falcon tube cap.
5. While holding the cap, apply N_2 gas at 8–10 psi for 2 min (*see* **Note 5**).
6. Remove the tubing and spin down the nebulizer 1 min at 1500*g* (*see* **Note 6**).
7. Precipitate the DNA with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
8. Keep at –20°C overnight.
9. Centrifuge at 12,000*g* for 30 min at 4°C.
10. Add 3 mL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
11. Dry in speedVac (*see* **Note 7**) and resuspend in the necessary vol of 5 m*M* Tris-HCl, pH 8.0, to reach a final vol of 100 μL after adding the reagents of the next step.
12. Transfer to a 1.5-mL tube and add 10 μL of dNTPs (0.5 m*M* each), 20 U T4 DNA polymerase, and 10 μL T4 DNA polymerase buffer.
13. Incubate 15 min at 30°C.
14. Add 6 U Klenow enzyme.

15. Incubate 15 min at 30°C.
16. Clean up through a QIAquick column (*see* **Note 8**).
17. Elute with 50 µL of 10 m*M* Tris-HCl, pH 8.0 (EB buffer; Qiagen).
18. Collecting the liquid in the same tube, re-elute with the necessary vol of water to reach a final vol of 100 µL after adding the reagents of the next step.
19. Add 5 U T4 PNK, 10 µL T4 PNK buffer, and 2 µL ATP 100 m*M*.
20. Incubate 30 min at 37°C.
21. Add 100 µL of water and extract with 200 µL of phenol:chloroform by vortex mixing and centrifuging at 12,000*g*.
22. Transfer the upper phase to a new tube and extract with 200 µL of chloroform by vortex mixing and centrifuging at 12,000*g*.
23. Transfer the upper phase to a new tube and precipitate with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
24. Leave at –20°C overnight.
25. Centrifuge at 12,000*g* for 30 min at 4°C.
26. Add 400 µL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
27. Dry and resuspend in 20 µL of 10 m*M* Tris-HCl, pH 8.0.

3.3. Adaptor Ligation

1. In a 1.5-mL tube, mix 10 µL of top adaptor oligonucleotide and 10 µL of bottom adaptor oligonucleotide (*see* **Note 9**).
2. Add 0.5 µL of 1 *M* NaCl.
3. Incubate 2 min at 75°C and anneal for at least 2 h by cooling down very slowly to 4°C.
4. In a new 1.5-mL tube, mix 10 µL of end-repaired DNA, 20 µL of annealed adaptor, 4 µL of T4 DNA ligase buffer, 10 U of T4 DNA ligase, and water to a final vol of 40 µL.
5. Incubate 24 h at 12°C (*see* **Note 10**).
6. Add 60 µL of TEN buffer (*see* **Note 11**).
7. Place the size fractionation column in a support and remove first the top and then the bottom cap (*see* **Note 12**).
8. Drain the liquid by gravity.
9. Wash the column by adding 800 µL of TEN buffer and allowing to drain completely.
10. Repeat the wash three more times.
11. Label 20 1.5-mL tubes and align them in a rack.
12. Add the adapted DNA to the upper frit of the column and allow to drain completely into the first 1.5-mL tube.
13. Add 100 µL of TEN buffer and collect the effluent in the second tube.
14. Add another 100 µL of TEN buffer and begin to collect a single drop per tube until complete drain.
15. Repeat the last step until 18 drops have been collected.
16. Run 3 µL of each fraction in an agarose gel.
17. Pool the first three fractions where DNA can be detected in the gel (*see* **Note 13**).

3.4. Vector Preparation

1. In a 1.5-mL tube, mix 2 µg of pUC 19 DNA, 30 U of *Xba*I, 6 µL of buffer H, and water up to 60 µL (*see* **Note 14**).
2. Incubate 2 h at 37°C.
3. Inactivate the enzyme incubating 20 min at 65°C.
4. Chill on ice and add 4 µL of buffer L, 2 µL of 10 mg/mL BSA, 4 µL of 1 m*M* dCTP, 8 U of Klenow enzyme, and water up to a final vol of 100 µL.
5. Incubate 30 min at 30°C.
6. Inactivate the enzyme incubating 15 min at 65°C.
7. Clean up the DNA through a QIAquick column.
8. Elute with 50 µL of 10 m*M* Tris-HCl, pH 8.0.
9. Re-elute in the same tube with 39 µL of water.
10. Add 10 µL of CIP buffer and 1 µL of 2 U/µL CIP.
11. Incubate 30 min at 37°C.
12. Add 2 µL 0.5 *M* EDTA and incubate 15 min at 65°C.
13. Add 100 µL water.
14. Extract with 200 µL of phenol:chloroform.
15. Extract with 200 µL of chloroform.
16. Precipitate with one-fiftieth vol of 5 *M* NaCl and 2 vol of ethanol.
17. Leave overnight at –20°C.
18. Centrifuge at 12,000*g* for 30 min at 4°C.
19. Add 500 µL of 70% ethanol and centrifuge at 12,000*g* for 10 min at 4°C.
20. Dry and resuspend in 100 µL of 10 m*M* Tris-HCl, pH 8.0 (*see* **Note 15**).

3.5. Preparation of Electrocompetent JM107 or JM107MA2 Cells

This protocol was modified from the manual by Sambrook and Russell ***(21)*** (*see* **Note 16**).

1. Use one JM107 or JM107MA2 colony from a fresh plate to inoculate 3 mL of LB medium. Incubate at 37°C overnight with shaking.
2. Take 2 mL of the overnight culture to inoculate 500 mL of SOB medium without magnesium. Incubate at 37°C shaking at 250–300 rpm until reaching an OD_{550} of 0.6–0.7.
3. Chill the culture on ice for 20 min and transfer to two 250-mL centrifuge bottles. Centrifuge at 2500*g* at 4°C for 15 min.
4. Repeat the wash in 10% glycerol. Discard the supernatant and resuspend each pellet in 10 mL of chilled 10% glycerol.
5. Transfer to two 14-mL centrifuge tubes.
6. Centrifuge at 2500*g* at 4°C for 15 min.
7. Resuspend both pellets in a total of 2 mL of chilled 10% glycerol.
8. Transfer 100 to 200-µL aliquots of the cells suspension to chilled sterile 1.5-mL microcentrifuge tubes. Freeze the cells in liquid N_2 and store at –70°C (*see* **Note 17**).

3.6. Ligation

1. In a 1.5-mL tube, mix 5–10 ng of vector, 10–100 ng of adapted and size fractionated genomic DNA (**step 17** from **Subheading 3.3.**), 1 µL of ligation buffer, 1 U of ligase, and take to a final vol of 10 µL with water.
2. Incubate 16 h at 12°C.
3. Add 90 µL of 10 m*M* NaCl.
4. Clean up the reaction using a QIAquick column, eluting in 50 µL of 10 m*M* Tris-HCl, pH 8.0.

3.7. Electroporation

1. Thaw electrocompetent cells in ice.
2. Mix 30 µL of cells with 1–3 µL of cleaned up ligation reaction in a chilled 1.5-mL tube.
3. Transfer the mixture to a chilled 0.1-cm gap electroporation cuvette and electroporate at 1.8 kV. Immediately add 750 µL of SOC medium and transfer to a sterile 14-mL centrifuge tube.
4. Incubate cells at 37°C for 45 min with gentle shaking.
5. Plate aliquots of approx 200 µL of cells together with 50 µL IPTG and 50 µL X-gal in LB-ampicillin plates.
6. Incubate overnight at 37°C.

3.8. Checking the Average Library Insert Size by Colony PCR

1. In a 1.5-mL tube, mix 60 µL of 10× PCR buffer, 30 µL of 10 µ*M* M13/pUC sequencing (–40) primer, 30 µL of 10 µ*M* M13/pUC reverse sequencing (–24) primer, 12 µL of dNTP mixture, 6 µL of 5 U/µL *Taq* DNA polymerase, and 462 µL of water (*see* **Note 18**).
2. Transfer 20 µL of the mixture to each of 30 250-µL PCR tubes.
3. Using an automatic pipet set in 5 µL, pick one white colony into the first PCR tube and pipet up and down a few times.
4. Repeat the last step for the rest of the tubes using a new tip each time.
5. Put the tubes in a PCR machine under the following program: 5 min at 95°C, then 25 cycles of: 30 s at 95°C, 45 s at 55°C, 3 min 30 s at 72°C, 10 min at 72°C, then forever at 4°C.
6. Run 10 µL of each reaction in an agarose gel.
7. Estimate the average insert size taking into account that the PCR fragments include 30–60 bp of vector sequence in each end. The proportion of clones containing repetitive DNA can be estimated as well (*see* **Note 19**).

4. Notes

1. For all buffers and solutions all Milli-Q® water (Millipore) is used.
2. When possible, it is preferable to use a tissue with low plastid content (i.e., maize immature ears). This would reduce the chloroplast DNA contamination. If the methylation status of a certain kind of gene is known to change with development, it should be taken into account at the moment of choosing the tissue for preparing DNA.

3. The use of a Polytron can be omitted if the blender properly homogenizes the tissue. In the case of hard tissue like pine needles, the Polytron may be necessary.
4. If the amount of starting material is small, DNA fibers may not be formed after adding isopropanol. In this case, the DNA can be recovered by centrifugation at 12,000*g* for 30 min.
5. The nebulization time and pressure need to be calibrated. Aliquots of DNA can be taken at different nebulization times and checked in agarose gels. The optimal nebulization conditions should break down the DNA to fragments mainly between 1 and 4 kbp.
6. As nebulizers are not designed for centrifugation, a rotor must be adapted to hold them. For example, the Sorvall® GSA rotor (NEN® Life Science Products) can be used if the bottoms of the wells are cushioned with paper towels.
7. The pellet is often loose and hard to see. It is advisable not to remove all the 70% ethanol and dry it for a longer time in the SpeedVac.
8. If a phenol extraction followed by ethanol precipitation is performed instead of the column clean up, a very hard to dissolve pellet is formed.
9. After annealed, the adaptor looks like this:

```
5'(P)-TAGACGCCTCGAG-3'
      | | | | | | | | | |
      3'-TGCGGAGCTC-5'
```

10. The 3-nucleotide overhang adaptor works very well. However, if necessary, cloning efficiency can be improved by using a double adaptor method ***(22)***.
11. Instead of using a column, the DNA can be size-fractionated by agarose gel electrophoresis. In this case, fragments ranging from 1–4 kbp must be eluted from the gel. One disadvantage of this approach is that a melting step needs to be performed by heating, which may denature the adaptor whose shorter oligonucleotide is not covalently linked. Using high quality low melting point agarose like SeaPlaque GTG agarose (BioWhittaker Molecular Applications) and the QIAquick gel extraction kit allows to melt the agarose at room temperature, which helps to overcome the problem. Alternatively, the shorter oligonucleotide can be added to the vector ligation reaction to improve the ligation efficiency.
12. To avoid the formation of bubbles inside the column, it is advisable to use a needle to make a hole in the top cap before removing it.
13. Taking the first 3 to 4 fractions in which DNA can be observed in the agarose gel usually works well. The next fractions may contain unligated adaptors and small DNA fragments, although they are not visible in the sample loaded in the gel. If no or few small insert clones are detected after estimating the library insert size (*see* **Subheading 3.8.**), the inclusion of more elution fractions can be considered for future construction of filtered libraries.
14. pUC 19 and *Xba*I are used as an example. Other vectors and restriction enzymes can be used as well. However, the protocols must be adapted accordingly in terms of selective antibiotic, adaptor sequence, host strain requirements, etc.
15. Before using a vector for library construction, some controls must be performed

by *E. coli* transformation: *(i)* vector with no ligase; *(ii)* self-ligated vector; and *(iii)* vector ligated to a control insert. The first two controls should yield no or very few blue colonies only. The third one should yield no or very few blue colonies and a large number of white colonies. In this case, the control insert is made by annealing the longer oligonucleotide used to make the adaptor and another 13-mer oligonucleotide: 5'(P)-TAGCTCGAGGCGT-3'. When annealed it looks like this:

```
5'(P)-TAGACGCCTCGAG-3'
          | | | | | | | | |
3'-TGCGGAGCTCGAT-(P)5'
```

16. JM107 ***(23)*** and JM107MA2 ***(24)*** are shown as examples of filtering and unfiltering strains, respectively. Other strains can be used, e.g., DH5α-E (*mcrBC*$^+$) and DH10B (*mcrBC*$^-$), both of which are available as electrocompetent from Invitrogen. If commercial strains are used, the protocols should be adapted to any special requirements of a particular *E. coli* strain. However, among *mcrBC*$^+$ strains, variations in filtering efficiency has been observed ***(14)***. Thus, both the transformation and filtering efficiencies need to be considered when choosing the strain to approach a large-scale methylation filtration project.
17. After a batch of competent cells is prepared, it must be tested by transforming a known amount of supercoiled plasmid. Usually the transformation efficiency is $>1 \times 10^{10}$ colonies/μg of plasmid DNA. Also, cells must be tested for any plasmid contamination by doing an electroporation without DNA, which should yield no colonies in selective medium.
18. The amount of PCR mixture can be increased to compensate for pipeting errors and to include some useful PCR controls like a blue colony, vector DNA, a water control, single primer controls, etc. This is a robust PCR assay and any commercially available PCR reagents should work as well as any combination of M13 forward and reverse primers. Instead of using PCR, insert sizes can be checked by doing plasmid minipreps of white colonies and subsequent restriction enzyme digestion and agarose gel electrophoresis.
19. An easy way to estimate the number of clones containing repetitive DNA is to bind a number of clones to a hybridization membrane and hybridize it against total labeled genomic DNA. In this labeled sample, only the repetitive DNA will be present in high enough proportion to produce a hybridization signal. Low copy DNA will be too diluted to show any hybridization. In this way, the high copy DNA containing clones can be identified as hybridizing clones. The proportion of high vs low copy clones can be compared to that in a control unfiltered library to estimate the filtering efficiency of the cloning process. The unfiltered library is constructed simply by transforming the same ligation mixture used for the filtered library into a *mcrBC*$^-$ *E. coli* strain. The hybridization can be performed on one to a few hundred clones from each library by colony hybridization ***(21)***. For example, for maize, where 80–90% of the genome is composed of repetitive DNA, a 5- to 10-fold decrease in the proportion of repetitive clones is expected in

a filtered vs a control library. There may be some variations due to the frequent methylcytosine to thymine transition. This mutation occurs frequently in silent repetitive DNA that is not under selective pressure. For this reason, some decayed repeats can be recovered in filtered libraries. Sequencing and Basic Local Alignment Search Tool (BLAST) analysis ***(25)*** of a few hundred clones from each library is an independent way to estimate how well the technique is working.

References

1. Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) Life with 6000 genes. *Science* **274,** 546–567.
2. The *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018.
3. The *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.
4. Blattner, F. R. (1983) Biological frontiers. *Science* **222,** 719–720.
5. Putney, S. D., Herlihy, W. C., and Schimmel, P. (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302,** 718–721.
6. Adams, M. D., Kelley, J. M., Gocayne, J. D., et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656.
7. Bento Soares, M. and Bonaldo, M. F. (1998) Constructing and screening normalized cDNA libraries, in *Genome Analysis. A Laboratory Manual. Vol. 2. Detecting Genes.* (Birren, B., Green, E. D., Klapholz, S., Myers, R. M., and Roskams, J., eds.), CSH Laboratory Press, Cold Spring Harbor, NY, pp. 49–158.
8. Barakat, A., Matassi, G., and Bernardi, G. (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proc. Natl. Acad. Sci. USA* **95,** 10044–10049.
9. Chandler, V. L. and Hardeman, K. J. (1992) The *Mu* elements of *Zea mays*. *Adv. Genet*. **30,** 77–122.
10. Raizada, M. N., Nan, G. L., and Walbot, V. (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. *Plant Cell* **13,** 1587–1608.
11. Raleigh, E. A. and Wilson, G. (1986) *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. *Proc. Natl. Acad. Sci. USA* **83,** 9070–9074.
12. Dila, D., Sutherland, E., Moran, L., Slatko, B., and Raleigh, E. A. (1990) Genetic and sequence organization of the *mcrBC* locus of *Escherichia coli* K-12. *J. Bacteriol.* **172,** 4888–4900.
13. Sutherland, E., Coe, L., and Raleigh, E. A. (1992) McrBC: a multisubunit GTP-dependent restriction endonuclease. *J. Mol. Biol.* **225,** 327–348.
14. Rabinowicz, P. D., Schutz, K., Dedhia, N., et al. (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23,** 305–308.

15. Walker, E. L. and Panavas, T. (2001) Structural features and methylation patterns associated with paramutation at the *r1* locus of *Zea mays*. *Genetics* **159,** 1201–1215.
16. Walbot, V. and Warren, C. (1990) DNA methylation in the *Alcohol dehydrogenase-1* gene of maize. *Plant Mol. Biol.* **15,** 121–125.
17. Patterson, G. I., Thorpe, C. J., and Chandler, V. L. (1993) Paramutation, an allelic interaction, is associated with a stable and heritable reduction of transcription of the maize *b* regulatory gene. *Genetics* **135,** 881–894.
18. Povinelli, C. M. and Gibbs R. A. (1993) Large-scale sequencing library production: an adaptor-based strategy. *Anal. Biochem.* **210,** 16–26.
19. Kiss, T., Toth, M., and Solymosy, F. (1985) Plant small nuclear RNAs. Nucleolar U3 snRNA is present in plants: partial characterization. *Eur. J. Biochem.* **152,** 259–266.
20. Wagner, D. B., Furnier, G. R., Saghai-Maroof, M. A., Williams, S. M., Dancik, B. P., and Allard, R.W. (1987) Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. *Proc. Natl. Acad. Sci. USA* **84,** 2097–2100.
21. Sambrook, J. and Russell, D. W. (eds.) (2001) *Molecular Cloning. A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.
22. Andersson, B., Wentland, M. A., Ricafrente, J. Y., Liu, W., and Gibbs, R. A. (1996) A "double adaptor" method for improved shotgun library construction. *Anal. Biochem.* **236,** 107–113.
23. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**, 103–119.
24. Blumenthal, R. M., Gregory, S. A., and Cooperider, J. S. (1985) Cloning of a restriction-modification system from *Proteus vulgaris* and its use in analyzing a methylase-sensitive phenotype in *Escherichia coli*. *J. Bacteriol.* **164**, 501–509.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.

3

RescueMu Protocols for Maize Functional Genomics

Manish N. Raizada

Summary

RescueMu is a modified *Mu1* transposon transformed into maize to permit mutagenesis and subsequent recovery of mutant alleles by plasmid rescue. *RescueMu* elements insert late in the germline as well as in terminally dividing somatic (e.g., leaf) cells. Germinal insertions may result in a mutant phenotype, and *RescueMu* permits recovery of 5–25 kb of transposon-flanking genomic DNA without having to construct and screen genomic DNA libraries. Late somatic insertions of *RescueMu* do not result in a visible phenotype, but they are instead used to construct plasmid libraries of gene-enriched maize genomic DNA to facilitate the identification and sequencing of the euchromatic portion of the maize genome. This is because maize leaves contain abundant independent *RescueMu* somatic insertions, and 70–90% of these insertions occur preferentially into genes and not repetitive DNA. This chapter describes detailed protocols on how to obtain, generate, and use *RescueMu* for maize genomics, including resources developed by the Maize Gene Discovery Project (MGDP) consortium available online at ZmDB.

Key Words

Mutator, *RescueMu*, maize, genomics, transposon, genome survey sequence, plasmid rescue, techniques

1. Introduction

Mutator (*Mu*) is a large DNA transposon family in maize (*see* refs. ***1,2*** for reviews). Traditionally, *Mu* has been used to create novel mutants randomly in the search for new genes (forward mutagenesis) and to create saturating populations of transposon insertions useful for reverse-genetics screens. This is due to several factors: first, 70–90% of *Mu* elements insert into genes ***(3)***, not into the repetitive DNA fraction which constitutes >80% of the maize genome ***(4)***. Second, heritable *Mu* insertions occur late in germinal cells resulting in sibling progeny that carry independent insertions. *Mu* elements insert at a high fre-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

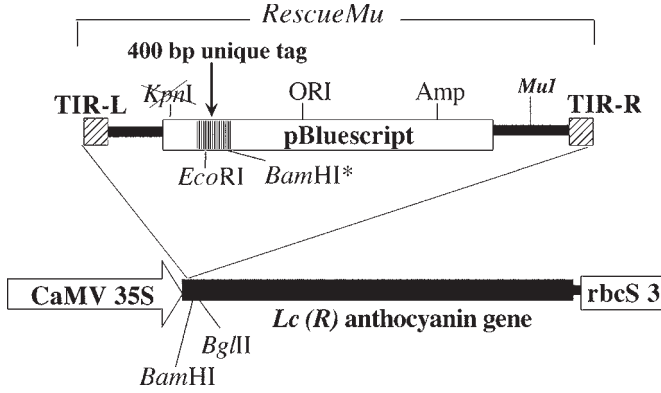


Fig. 1. Structure of the *RescueMu* vector. *RescueMu* consists of a plasmid inserted into an intact *Mu1* nonautonomous element. *RescueMu* is inserted downstream of a CaMV 35S promoter in the 5' untranslated leader of maize *Lc (Leaf Color)* a transcription factor of the *R* family required for anthocyanin production. Excision of *RescueMu* can restore tissue pigmentation. Two elements, *RescueMu2* and *RescueMu3,* differ by the presence of unique 400 bp heterologous tags of *Rhizobium* DNA, and both are present in the original *RescueMu* transgenic lines. The asterisk indicates that the internal *Bam*HI site is present in *RescueMu3*, but absent in *RescueMu2*.

quency (10^{-6} – 10^{-4} per locus per generation), to both linked and unlinked loci where they remain stable and transmissible through the germline. A mutant caused by a *Mu* element rarely ever reverts to wild-type. In contrast, maize *Ac/Ds* elements and *En/Spm* elements insert stochastically during maize development, preferentially insert within a 5 cM region of the donor site and may excise in subsequent generations (reviewed in ref. ***1***). Finally, because inherited *Mu* elements are not lost and continue to duplicate, they amplify over generations, up to hundreds of copies per plant, unlike *Ac/Ds* transposons that are inhibited by a negative feedback transposition control mechanism. Thus, random gene-targeted *Mu* amplification permits saturation mutagenesis.

Each member of the *Mu* element family is defined as sharing a common approx 215 bp terminal inverted repeat (TIR) to which the *Mu* transposase binds (reviewed in ref. ***1***). *MuDR* is a 4.9-kb *Mu* element that encodes two proteins required for transposition. The *Mutator* family was likely created by internal deletion and recombination of *MuDR* resulting in at least eight non-protein-coding subfamilies of smaller transposons (*Mu1–Mu8*), which are incapable of autonomous transposition, but may transpose in the presence of a functional *MuDR* element.

*RescueMu*2 and *RescueMu3* (**Fig. 1**) are modified *Mu1* elements into which high-copy number bacterial plasmids conferring ampicillin resistance were

inserted (3). They differ only by the presence of an internal 400-bp sequence tag derived from *Rhizobium*. These plasmids were stably co-transformed with the pAHC20 plasmid into maize by biolistic transformation. pAHC20 is a plasmid encoding *bar*, which is a selectable marker gene that confers resistance to the herbicide glufosinate/Basta ***(5)***. *RescueMu* transgenic lines must be crossed to an active *MuDR* line to transpose ***(3)***.

RescueMu was constructed to accelerate the discovery and characterization of *Mu*-mutagenized genes underlying mutant phenotypes of interest. Plasmid rescue can now be used to recover 5–20 kb of *Mu* element flanking DNA in plasmid form ready for DNA sequencing in only a few days ***(3)***, instead of having to construct a genomic library from a mutant plant.

In addition to germinal insertions, research using *RescueMu* uncovered that *Mu* elements also transpose at a very high frequency in terminally dividing somatic cells (e.g., leaf cells) ***(3)***. Late somatic *RescueMu/Mu* insertions are unlikely to cause a noticeable phenotype, and because they rarely occur in the shoot apical meristem, they are usually not transmitted to the next generation. However, the somatic behavior of *RescueMu* has created a novel resource for the construction of bacterial libraries of euchromatic-rich maize genomic DNA in plasmid form ready for DNA sequencing. This is because *RescueMu* somatic insertions also occur preferentially into genes ***(3)***. Read-out DNA sequencing from *RescueMu* elements recovered from a single leaf can rapidly identify significant numbers of independent genes and gene-rich DNA sequence ***(3)***. Because of the sensitivity of bacterial transformation and antibiotic selection, *RescueMu* insertions contained in single small leaf sectors can be recovered in *Escherichia coli* from a pool of plant material, filtering out all other maize genomic DNA. These features permit *RescueMu* sequencing to be an alternative to expressed sequence tag (EST) sequencing for gene discovery while offering several unique advantages: unlike EST sequencing, *RescueMu* may be used to find poorly transcribed genes. Second, *RescueMu* may lead to the discovery of large numbers of nontranscribed regulatory regions in maize located near *RescueMu* insertions ***(3)***, something not possible by EST sequencing. Finally, *RescueMu* sequencing from both the right and left borders allows more transcribed sequence to be obtained, including complete 5' and 3' untranslated regions. Whereas *RescueMu* plasmids can include up to 25 kb of genomic DNA ***(3)***, alternative methods to isolate genomic DNA flanking *Mu* insertions such as thermal asymmetric interlaced polymerase chain reaction (TAIL-PCR) using *Mu* read-out primers ***(6,7)*** typically result in <500 bp of readable DNA sequence.

The Maize Gene Discovery Project (MGDP) is a consortium of laboratories headed by Virginia Walbot (Stanford University) that is employing *RescueMu* on a large scale to accelerate the recovery of mutant-causing germinal

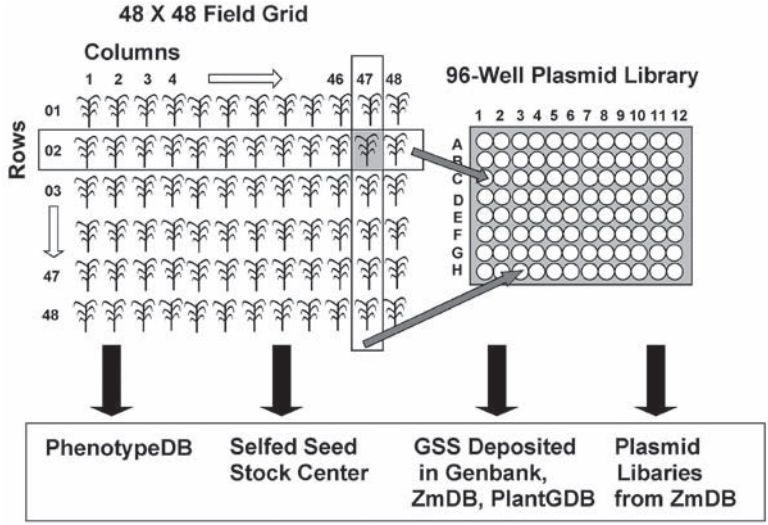


Fig. 2. Summary of *RescueMu* materials available from the MGDP.

RescueMu insertions and to construct libraries of *RescueMu*-mutagenized leaf DNA for maize euchromatic DNA sequencing. The MGDP makes available populations of *RescueMu* mutagenized seed, online descriptions of mutants, and 96-well microtiter plate libraries of recovered *RescueMu* plasmids representing somatic and germinal insertions. Each plate library represents plasmids recovered from a field grid consisting of 48 rows and 48 columns (2304 *RescueMu* plants) (**Fig. 2**). Each well contains *RescueMu* plasmids recovered from one row or one column (48 plants) in the grid. Each plant in the row or column is sampled by taking leaf punches from a single leaf. However, each plant is sampled twice, one leaf for the row sample and the second leaf for the column sample. If a *RescueMu*-flanking genomic DNA sequence is recovered in both a row and a column of a grid, the logical intersection identifies the single plant in the grid as the donor of the common *RescueMu* allele. Because each row and column are sampled from separate leaves, and because only a germinal insertion would be expected to extend beyond a single leaf, then double-sampling is used to distinguish between the more frequent late somatic insertions (leaf sector) and the rarer germinal insertions (whole plant). The MGDP makes available approx 100–500 bp read-out sequences from these libraries, known as genome sequence surveys (GSSs), which may be queried online at GenBank®, PlantGDB, or ZmDB. For online links, detailed information, or to order materials, the reader is encouraged to visit the Web site of the MGDP, known as ZmDB (www.zmdb.iastate.edu).

The first part of this chapter describes how to generate, recover, and analyze novel *RescueMu* insertions in-house, including: *(i)* how to obtain and choose *RescueMu* seed stocks; *(ii)* how to perform *RescueMu* plasmid rescues from maize; *(iii)* how to select against contaminating plasmids using restriction enzymes and filter hybridization techniques; and *(iv)* how to read-out and analyze sequence from recovered *RescueMu* elements. In **Subheadings 2.8.** and **3.8.**, I have included additional descriptions on how to request and use materials generated by the MGDP in combination with these basic protocols.

2. Materials

2.1. Selecting RescueMu *Plant Material to Generate Novel Insertions*

1. Glufosinate ammonium/phosphinothricin-tripeptide (PPT)/Basta (Liberty® Herbicide; Aventis Crop Science).
2. Tween® 20.

*2.2. Genomic DNA Isolation (*see *Note 4)*

1. Chloroform.
2. Isoamyl alcohol.
3. Isopropanol.
4. 70% (v/v) Ethanol.
5. Water.
6. TE buffer: 10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA.
7. Prepare plasmid-free CTAB buffer: 100 m*M* Tris-HCl, pH 7.5/8, 2% (w/v) CTAB, 1.4 *M* NaCl, 20 m*M* ethylene diamine tetraacetic acid (EDTA), pH 7.5/8, 1% (v/v) β-mercaptoethanol, 1% (w/v) sodium bisulfite. For 100 mL of CTAB buffer, dissolve CTAB in 60 mL water by heating in a microwave for 20 s and then add other components. Add β-mercaptoethanol just before use. Store at room temperature or 4°C.

2.3. Plasmid Rescue

1. Enzymes needed: *Kpn*I, RNaseA, *Bgl*II, *Eco*RI, T4 DNA ligase (Invitrogen).
2. ElectroMAX DH10B competent cells (>10^{10} colony-forming units [cfu]/μg) (Invitrogen or LIFE Technologies).
3. 3 *M* Sodium acetate.
4. Buffer-saturated phenol, pH 8.0.
5. Chloroform.
6. Isoamyl alcohol.
7. 70% (v/v) Ethanol.
8. Water (plasmid-free).
9. SOC media (Invitrogen or LIFE Technologies).
10. DNA Electroporator and 0.1-cm cuvettes.
11. LB-carbenicillin (100 mg/L) Petri plates (*see* **Note 9**).

2.4. Isolating DNA Fragments for DNA Hybridization Probing of Rescued Colonies

1. Enzymes needed: *Pst*I, *Sac*I, *Xba*I, *Xho*I, *Bsp*HI.
2. Plasmids needed: pR, pBluescript® KS (Stratagene), pRescueMu2 and pRescueMu3, pMR15 and pMR17 (*see* **Note 13**).
3. *RescueMu* probe amplification primers: *(i)* primer p173+155F GCGAATTC GACAGCCGGCAGGGCATTC; *(ii)* T7 primer CGCGTAATACGACTCACT ATAGGGC; and *(iii)* primer p192+130F TTCCTGCAGCGGCCGCGGATCAG.

2.5. Preparing Filters for Screening of Rescued Colonies

1. Whatman 3 MM filter paper (Whatman).
2. 0.5 *M* NaOH.
3. 1 *M* Tris-HCl (pH 7.5).
4. UV cross-linker (e.g., Stratalinker®; Stratagene).
5. India ink.
6. Nitrocellulose filters (e.g., NEN Colony/Plaque Screen).
7. 80°C Oven (if using nitrocellulose).

2.6. Confirming RescueMu Insertions Using Colony-Lift Hybridizations

1. Random primer labeling kit (e.g., DecaPrimeII; Promega).
2. ^{32}P-α [dCTP] (2000–3000 Ci/mmol) (Amersham Pharmacia Biotech).
3. NucTrap Push Columns (Stratagene).
4. 2× SSC, pH 7.0: 0.3 *M* sodium citrate, 0.3 *M* NaCl.
5. 10% (w/v) Sodium dodecyl sulfate (SDS).
6. 10 mg/mL Salmon sperm DNA.
7. Prehyb buffer: 1% (w/v) SDS, 2× SSC, 10% (w/v) dextran sulfate, 50% deionized formamide, 3× Denhardt's reagent (1% [w/v] Ficoll® 400, 1% [w/v] polyvinylpyrrolidone, 1% [w/v] bovine serum albumin [Fraction V; Sigma]).

2.7. Analyzing and Sequencing of RescueMu Plasmids

1. Enzymes needed: *Kpn*I, *Hin*dIII, *Eco*RI.
2. Sequencing primers: *Mu*3-R TGCTGTCTTGTGTCCGTTTTA and *Mu*3-L AGCTGTCTCGTATCCGTTTTG.

2.8. Requesting RescueMu MGDP Materials

1. 96-Well plates of *RescueMu* plasmids, each recovered from a field grid of 48 × 48 plants, may be purchased for $150 US at (www.zmdb.iastate.edu). Click on Order Materials, then follow the Library Plate link.
2. Pictures and descriptions of visible mutants in each MGDP *RescueMu* field grid may be found at the ZmDB Maize Phenotype Database (PhenotypeDB) at (http://www.zmdb.iastate.edu/zmdb/phenotypeDB/index.htm). Selfed seed from these grid plants are available from the Maize Genetics Cooperation-Stock Center (http://w3.ag.uiuc.edu/maize-coop/mgc-home.html) (*see* **Note 1**). Send an e-mail

to maize@uiuc.edu indicating the *RescueMu* field grid letter, row, and column numbers.

3. To screen *RescueMu* 96-well plasmid libraries by PCR to search for an insertion into sequence of interest (reverse genetics), the right-side *RescueMu* read-out primer (*Mu1–R*) is 5'-TAT TTC GTC GAA TCC GCT TCT-3', and the left-side read-out primer (*Mu1–L*) is 5'-CAT TTC GTC GAA TCC CCT TCC-3'.

3. Methods

3.1. Selecting RescueMu Plant Material to Generate Novel Insertions

1. Request and select active *RescueMu* seeds (*see* **Notes 1–3**).
2. Confirm the presence of the *RescueMu* transgene via its linkage to plasmid pAHC20 ***(5)***, which encodes resistance to the herbicide glufosinate/PPT/Basta. To test for herbicide resistance, a 5-cm-diameter circle is made using a black marker onto a leaf, which is then painted with 0.75% (v/v) glufosinate ammonium (Liberty Herbicide, 18% [v/v] solution) containing 0.1% (v/v) Tween 20 using a Q-tip. Only plants that are non-necrotic, 5–7 d after herbicide application, should be used.

3.2. Genomic DNA Isolation

1. Using plasmid-free solutions (*see* **Note 4**), isolate genomic DNA, preferably from young leaves 1–5, using the urea extraction method ***(8)*** or the CTAB method below ***(9)***. Both methods work well.
2. Grind 0.1–0.3 g of tissue to a fine powder in liquid nitrogen using a mortar and pestle.
3. Add tissue to a 2-mL Eppendorf® tube containing 0.9 mL of CTAB buffer.
4. Vortex mix sample briefly and keep on ice until all samples are ground.
5. Incubate the tubes at 60°C for 30 min, then cool at room temperature 10 min.
6. Add 1 vol chloroform:isoamyl alcohol (24:1) and invert tubes continuously for 5 min.
7. Centrifuge the tubes 5 min in a microcentrifuge at >14,000*g*, then remove the upper aqueous phase to a clean 2-mL Eppendorf tube.
8. Repeat **steps 6** and **7**. Transfer the upper, aqueous phase to a 1.5-mL Eppendorf tube.
9. Add 1 vol isopropanol and invert the tubes gently until the DNA precipitates.
10. Either spool the DNA with the curled-by-flaming tip of a sterile Pasteur pipet or minicentrifuge for 2 min at >14,000*g*.
11. Resuspend the DNA in 1 mL of 70% (v/v) ethanol. Incubate at room temperature 20 min.
12. Centrifuge the tube at >14,000*g* for 15 s, then air-dry the pellet.
13. Dissolve the DNA in 50–200 μL of TE. Incubate at 4°C to dissolve.
14. Store at –20°C until next step.

3.3. Plasmid Rescue

1. Digest 10 μg of genomic DNA with 50 U of *Kpn*I in the presence of RNaseA in a vol of 150 μL, for 90 min at 37°C (*see* **Notes 5** and **11**).
2. Add 150 μL of phenol:chloroform:isoamyl alcohol (25:24:1), mix by inversion, microcentrifuge at >14,000*g*, remove the upper aqueous phase to a fresh tube. Repeat once (*see* **Note 6**).
3. Add 100 μL of chloroform, mix by inversion, microcentrifuge at >14,000*g*, and remove upper aqueous phase to a fresh tube.
4. To precipitate the DNA, add one-tenth vol of 3 *M* sodium acetate, mix by tapping, then add 2.5 vol of 95% ethanol.
5. Centrifuge for 20 min at >14,000*g* at 4°C.
6. Wash the pellet with 1 vol 70% (v/v) ethanol, then air-dry.
7. Dissolve in >20 μL water.
8. An optional *Bgl*II selection step (*see* **Note 3**) is performed as follows: digest DNA with 30 U of *Bgl*II in a final vol of 100 μL for 1 h at 37°C. Extract once with 1 vol of phenol:chloroform:isoamyl alcohol (25:24:1), then once with 1 vol chloroform as in **steps 2** and **3**. Ethanol precipitate and wash with 70% (v/v) ethanol as in **step 4**, but dissolve the final DNA pellet in >50 μL water.
9. Self-ligate at 14°C for 16 h with 10 U of T4 DNA ligase and 100 μL of fresh 5× ligation buffer (Invitrogen or LIFE Technologies) in a final vol of 500 μL (*see* **Notes 7** and **10**).
10. Extract the ligation mixture twice with 500 μL of phenol:chloroform:isoamyl alcohol (25:24:1) and once with 500 μL of chloroform as in **steps 2** and **3**.
11. Precipitate the DNA by adding one-tenth vol of 3 *M* sodium acetate, mix by tapping, then add 1 vol isopropanol. Invert.
12. Centrifuge 20 min, 14,000*g*, 4°C. Wash the pellet with 500 μL of 70% (v/v) ethanol and air-dry.
13. Dissolve the pellet in 10 μL water.
14. For each sample, aliquot 1 mL of SOC medium in a 3 to 10-mL tube.
15. For electroporation, thaw 30–50 μL of ElectroMAX DH10B cells (>10^{10} cfu/μg DNA) in an ice slurry exactly according to the manufacturer's recommendations (*see* **Notes 8** and **10**).
16. As the cells are thawing, aliquot 2 μL of DNA (approx 1 μg) per sample in a separate Eppendorf tube and chill on ice (*see* **Note 10**).
17. When the cells are thawed, aliquot 30–50 μL of cells in each tube containing the DNA and incubate on ice >1 min.
18. Just prior to each electroporation, pipet up the SOC media in a Pasteur pipet, ready for pipeting into the cuvette immediately after electroporation. A delay of only 20–30 s in the addition of SOC causes a significant decrease in transformation efficiency.
19. Electroporate exactly according to the instructions accompanying the competent cells. For a Bio-Rad device, cells are placed in a 0.1-cm gap disposal cuvette (Bio-Rad) set at 100 ohms, 2.5 kV, 25 μF, then discharged (time constant approx 2.3).

20. Immediately add 1 mL of SOC media into the cuvette, pipet up and down gently once, then remove into the 3 to 10-mL tube.
21. Shake at 37°C for 1 h at 225–300 rpm to allow expression of the antibiotic resistance gene.
22. To concentrate the cells, aliquot the SOC bacterial media into a 1.5-mL Eppendorf tube, and microcentrifuge for 5 s at 14,000*g* at room temperature.
23. Remove the SOC and gently resuspend in 200 μL of fresh SOC.
24. Plate 20 and 180 μL of cells onto ampicillin–carbenicillin-containing LB plates (*see* **Notes 9** and **11**).

3.4. Isolating DNA Fragments for DNA Hybridization Probing of Rescued Colonies (see Note 12)

1. The *RescueMu2*-specific probe is obtained as a 520-bp *Xho*I-*Xba*I fragment from pMR15 (*see* **Note 13**).
2. The *RescueMu3*-specific probe is obtained as a 478-bp *Xho*I-*Sac*I fragment from pMR17.
3. Alternatively, PCR may be used to amplify *RescueMu*2 and *RescueMu3* probes. To amplify *RescueMu2*, use 5' primer p173+155F and the 3' T7 primer. To amplify *RescueMu3*, use the 5' primer p192+130F and the 3' T7 primer 3. PCR cycle conditions are 94°C for 45 s, 50°C for 45 s, and 72°C for 60 s (30–35 cycles) in the presence of 2 m*M* $MgCl_2$. PCR products should be purified on an agarose gel.
4. Instead of using *RescueMu*-specific probes to detect new *RescueMu* insertions, an ampicillin probe may also be used. It is isolated as a 1-kb *Bsp*HI fragment from pBluescript KS+ and will detect both *RescueMu* plasmids.
5. Cauliflower mosaic virus (CaMV) 35S and maize *R(Lc)* probes should also be isolated to be used to screen against the recovery of the original *Lc*::*RescueMu* alleles after plasmid rescue (*see* **Note 12**). The CaMV 35S probe extends from +7072 to +7565 ***(10)*** and is isolated as a *Xba*I-*Pst*I fragment from plasmid pR ***(11)***. The maize *R(Lc)* probe is isolated as an approx 800-bp *Pst*I fragment from pR (*see* **Note 13**).

3.5. Preparing Filters for Screening of Rescued Colonies

1. This is the Grunstein-Hogness method ***(12)***.
2. Chill bacterial plates at 4°C for >1 h.
3. Lay out 4 pieces of cellophane (each >15 × 15 cm). Label 1, 2, 3, and 4. Place a square of Whatmann 3 MM blotting paper (>10 × 10 cm) beside, though not touching, each piece of cellophane. Have a timer ready for each of the four stations. Have a bottle of India ink with a gauge needle ready.
4. Pipet 1 mL of 0.5 *M* NaOH onto each of cellophane 1 and 2, and 1 mL of 1 *M* Tris-HCl buffer (pH 7.5) onto each of cellophane 3 and 4.
5. Use forceps to place a dry piece of nitrocellulose membrane onto each bacterial plate, one at a time. Wait 2 to 3 min. During this time, use a unique dot pattern and stab the membrane and LB with the India ink. This will be used to orient the

X-ray film after hybridization with the bacterial plates to pick positive *RescueMu* clones.

6. Transfer the filter onto cellophane 1 directly onto the pool of NaOH, colony-side facing up. Incubate 2 min to lyse the cells.
7. Transfer onto cellophane 2 and again incubate for 2 min as in **step 6**. Briefly blot onto Whatmann 3 MM paper to remove excess NaOH.
8. Transfer filter onto cellophane 3, directly onto solution of 1 M Tris-HCl, colony-side up. Incubate 2 min. Briefly blot onto Whatmann paper.
9. Transfer onto cellophane 4 and repeat as in **step 8**. Blot onto Whatmann 3 MM paper.
10. Immobilize DNA by UV cross-linking using manufacturer's recommendations, then place in an 80°C oven for 2 h. Store in a dry place until needed. Store the LB plates at 4°C.

3.6. Confirming RescueMu *Insertions Using Colony-Lift Hybridizations*

1. Prepare 10–50 ng of radioactive probe DNA using a random prime labeling kit (e.g., DecaPrimeII) and ^{32}P-α [dCTP]. Incubate at 37°C for >3 h, and then purify on a NucTrap push column to remove unincorporated nucleotides.
2. In the first round of hybridization, to identify plasmid contamination (*see* **Note 4**), colonies should be hybridized to a mixture of the two *RescueMu*-specific probes (*See* **Subheading 3.4.**) to confirm colony identity as described below.
3. Filters should be wetted in 2× SSC for 1 min, then prehybridized in Prehyb buffer in the presence of 0.1 mg/mL single-stranded DNA (prepared by boiling a 10 mg/mL stock of salmon sperm DNA for 5 min, then quick-chilled on ice). The filters should be incubated for 30 min to 24 h at 42°C in a shaking tupperware container or hybridization oven.
4. Following prehybridization, radiolabeled probe should be denatured by boiling for 5 min with 50% (v/v) formamide, then quick-chilled on ice. The denatured probe should be added directly to the filters in Prehyb buffer, and hybridization carried out for 16–24 h at 42°C.
5. The hybridization solution should be removed and the filters washed in 0.2× SSC/0.1% (v/v) SDS at 65°C (100–500 mL/10 filters) for 15 min, with 2 changes of wash buffer. The filters should be wrapped in cellophane paper and exposed to X-ray film for 6–24 h.
6. Using the India ink markings on the filters, the X-rays should be marked, allowing them to be aligned with each original LB plate.
7. Positive colonies from the first hybridization screen should be picked with sterile toothpicks, arrayed on duplicate LB plates (50–100/100-mm-diameter LB plate) and numbered. The plates are then incubated overnight at 37°C.
8. This entire procedure (**steps 1–7**) should be repeated on the duplicate plates of selected positive colonies in order to screen out colonies that represent recovery of the original *RescueMu*/pAHC20 transgene array (*see* **Note 12**). Colonies from the first plate should be hybridized to a mixture of CaMV 35S- and maize *Lc*(*R*)-specific probes; colonies from the second plate should be hybridized again to the

mixture of *RescueMu*-specific probes. Colonies that are positive for the *RescueMu* probes but negative for CaMV 35S and *Lc(R)*, should then be chosen for DNA sequencing.

3.7. Analyzing and Sequencing of RescueMu *Plasmids*

1. As a final check to confirm that the selected colonies represent true *RescueMu* insertions, plasmid DNA should be isolated and digested with *Kpn*I and *Hin*dIII. If a plasmid corresponds to a new insertion, there should be at least one fragment >4.7 kb (*see* **Note 4**). A comparison of restriction patterns of plasmids recovered from the same plant may be useful in determining if the recovered plasmid represents a somatic or germinal insertion (*see* **Note 14**).
2. For cleaner sequencing of flanking genomic DNA, plasmids may first be linearized with *Eco*RI, then repurified by ethanol precipitation.
3. For sequencing, the primers are located –122 bp from the outside edge of *RescueMu*. The right TIR out primer is *Mu3*-R and the left border TIR out primer is *Mu3*-L.
4. The first several bases will correspond to *Mu1* TIR sequence, followed by novel sequence. The first 9 bp immediately flanking TIR sequence should be duplicated at both the left and right borders of *RescueMu*, which is a hallmark of *Mu*/*RescueMu* transposition.

3.8. Using Existing RescueMu *MGDP Resources*

3.8.1. How to Query MGDP RescueMu Plasmid Library GSS Databases

1. The *RescueMu* GSS collection consists of tens of thousands of partial read-out sequences from recovered *RescueMu* elements, representing both somatic and germinal insertions in pools of maize leaves (**Fig. 2**) (*see* **Note 15**).
2. Go to (www.zmdb.iastate.edu) and click on the Search ZmDB button.
3. To search for a sequence of interest in the GSS collection, use the ZmDB Basic Local Alignment Search Tool (BLAST). In the new page, specify GSS database, enter the sequence, and then Run BLAST.
4. A ZMDB BLAST Results page will open to indicate if a successful alignment was found.
5. In the Results Summary box, look for the word *RescueMu* under Description. Click on the corresponding sequence name; this will open up a new page.
6. At the bottom of the new page, there will be a box to indicate if the *RescueMu* GSS aligns with maize EST sequences. There will be a second box that indicates which field grid library the GSS was obtained from (e.g., Library 1006 Grid G) and the plant location within the grid (e.g., row 16).
7. Alternatively, *RescueMu* GSS sequences may also be accessed using GenBank® National Center for Biotechnology Information (NCBI) by delimiting the search to the dbGSS database or via the Plant Genome Database at (www.plantgdb.org). PlantGDB permits other useful search options such as searching using a text identifier:

a. In PlantGDB, specify GSS or GSS contig under Sequence and *Zea mays*.
b. A query results page will open and list any *RescueMu* sequences that match the text.
c. Clicking on a sequence name will open up a new page that will indicate if the *RescueMu* GSS is part of a larger GSS contig and/or aligns with maize ESTs.

8. To identify upstream and downstream sequences to the original query sequence, look for overlapping ESTs or GSS contigs. For example, if RMTuc appears in the Results page, click on the corresponding link. This will open up a new page specifying that the GSS is part of a *RescueMu* tentative unique contig (RMTUC) assembled by aligning overlapping *RescueMu* GSSs and displaying overlapping EST sequences. For an example, go to (www.plantgdb.org), select Text Search, type in myb, specify GSS contig and *Zea mays,* then hit Search.

3.8.2. How to Retrieve a RescueMu *Genomic Plasmid from a MGDP Grid Library for Further Sequencing*

1. This section describes how to retrieve a plasmid encoding a GSS of interest from a MGDP 96–well grid library of *RescueMu* recovered plasmids in order to sequence further upstream or downstream. *RescueMu* GSS plasmids are not individually distributed by the MGDP.
2. Perform a sequence similarity search against the *RescueMu* GSS collection (*see* **Subheading 3.8.1., steps 1–6**).
3. In the Query Results page, note the grid origin of the GSS sequence (e.g., Grid G).
4. To identify the precise location of the GSS in a 96-well plate and the direction of the read-out sequence, locate the sequence identification (I.D.). Examples are 1006162C04.x2 1006 and 1008035A02.y1 1008 (*see* **Note 15**).
5. Purchase the correct 96-well *RescueMu* grid library plate online (*see* **Subheading 2.8.**).
6. After receiving the plasmid library, there are two methods to retrieve the GSS plasmid of interest from the correct well, PCR, or bacterial colony hybridization.
7. To PCR amplify the entire maize genomic DNA insert flanking *RescueMu* (up to 25 kb):
   a. Design a PCR primer specific to the GSS to amplify in the direction away from *RescueMu*.
   b. Synthesize a *RescueMu* read-out primer located on the opposite edge of the genomic insert. For example, if the GSS is from an "x" (right TIR) sequence, then the *RescueMu* read-out primer should correspond to the left TIR. The *RescueMu* left primer is 5'-CACCGCCGTGCTGCCGTAGAGCG-3' and the *RescueMu* right primer is 5'-CGCGTGACTGAGATGCGACGGAG-3'. These are located >220 bp internal to the left or right edge of the *RescueMu* element.
   c. Use MasterAmp Extra Long DNA Polymerase with High Fidelity 2× Extra Long PCR Premix 9 (Epicentre), 5 ng of library plate DNA, the GSS primer, and the *RescueMu* primer.
   d. Following an initial denaturation at 94°C for 1 min, perform 40 PCR cycles as

follows: 94°C for 15 s, 60°C for 30 s, and 68°C for 25 min. The long extension time is to amplify inserts up to 25 kb in length.

e. Additional details for PCR amplification may be found at (www.zmdb.iastate.edu); click on the Protocols button and follow the PCR link.

8. For bacterial colony hybridization, transform the DNA from the correct well (e.g., C04 or A02) into *E. coli* strain DH10B and ensure that the colonies are well separated (*see* **Subheading 3.3., steps 14–24**).
   a. To screen colonies containing the GSS plasmid of interest, generate a DNA probe corresponding to the GSS by PCR using the library well DNA or maize genomic DNA as the template. Alternatively, request an overlapping EST fragment (available online from ZmDB) to use as probe.
   b. Follow **Subheadings 3.5.** and **3.6.** to immobilize the bacterial colonies onto nylon–nitrocellulose and to screen colonies using the radiolabeled probe.
   c. Isolate plasmid DNA from positive colonies.
9. Confirm the identity of the recovered clone by DNA sequencing, and then design specific DNA sequencing primers to sequence upstream and downstream of the GSS.
10. For PCR cycle sequencing, consult (www.zmdb.iastate.edu); click on the Protocols button and follow the Cycle Sequencing link.

3.8.3. How to Use an EST or Heterologous Sequence to Screen RescueMu *Libraries by Reverse Genetics*

1. This section describes how to use a sequence (EST, heterologous sequence) with no similarity in the online ZmDB GSS collection to screen *RescueMu* plasmid libraries generated by the MGDP to identify a somatic or germinal insertion by reverse-genetics.
2. Purchase 96-well *RescueMu* grid library plates online.
3. Synthesize the *RescueMu* read-out primers (*Mu1*-L and *Mu1*-R) (*see* **Subheading 2.8.**).
4. Design and synthesize two or more PCR primers for the sequence of interest, both 5' to 3', one for the top strand and one for the bottom strand.
5. Perform a 96-sample PCR using the 4 PCR primers and use the following initial conditions: 0.5 m*M* dNTPs, 2.5 m*M* Mg^{++}, 0.8 μ*M* of each specific primer, 4.0 μM of each *RescueMu* primer, 2 U *Taq* DNA polymerase and 5 ng library plate DNA. Denature 95°C for 5 min, then amplify for 40 cycles (95°C for 30 s, 55°C for 30 s, then 72°C for 2 min), followed by a single extension at 72°C for 5 min.
6. Consult (www.zmdb.iastate.edu) for a grid-specific list of positive control PCR primers and other recommendations. Click on RMu Libraries and then Screening.
7. Sequence the fragment to confirm its identity.
8. If the insertion of interest is found in both a row well and a column well, this indicates a likely germinal insertion event and pinpoints the exact plant. Note the grid letter, row and column number to request seed from the Maize Genetics Cooperative-Stock Center (*see* **Note 1**).

3.8.4. How to Screen the MGDP RescueMu PhenotypeDB to Obtain a Mutant of Interest

1. Grids of 48 × 48 *RescueMu* plants have been screened by the MGDP for visible mutant phenotypes and descriptions are available online (*see* **Subheading 2.8.**). Mutants may be caused by either *RescueMu*, but more likely by background *Mu/MuDR* elements. Go to the PhenotypeDB index page at (www.zmdb.iastate.edu/zmdb/phenotypeDB/index.htm).
2. For relative mutation frequencies in each grid, consult the Grid Summary Table.
3. Begin the search by taking the Interactive PhenotypeDB Tutorial.
4. Choose one of three search tools. To search using a specific phenotype, for example a Knotted adult leaf, then use the Phenotype Lists search engine. To search by general category, for example all adult leaf mutants, then use the Mutant Browser. To search by a specific location within a grid, use the Location Search engine.
5. Hit Start Search.
6. In the Query Results page, the column and row of each mutant is listed. Click on the corresponding Grid letter; this opens up the PhenotypeDB Search Details page, which is a summary card of the scoring details.
7. At the bottom of the PhenotypeDB Search Details page, there are links to all the *RescueMu* GSSs recovered from the row and column pool that contained the mutant plant.
8. Use the grid, row, and column information to request selfed seed from the Maize Genetics Cooperation Stock Center (*see* **Subheading 2.8.** and **Note 1**).
9. Once seed have been received, the user may wish to backcross to create an isogenic background. *RescueMu* seed populations are in a mixed genotype, typically A188 > W23 > Robertson > K55 > Freeling > B73. For more details, go to (www.zmdb.iastate.edu), open the *RescueMu* Index menu on the right side and choose *RescueMu* Tagging Populations.

3.8.5. How to Use a RescueMu GSS to Identify a Corresponding Mutant Phenotype

1. Perform a BLAST search in ZmDB. Select the GSS database (*see* **Subheading 3.8.1.**).
2. In the Results Summary page, note whether the GSS appears as a single hit or multiple hits in the same library grid (indicated by the first 4 or last 4 letters under Description). Determine the row or column source of each GSS (see **Subheading 3.8.1., steps 3–6**).
3. If the GSS appears as only a single hit within any one grid, then proceed with **step 3**. For multiple hits, go to **step 4**.
   a. There is a high probability that the *RescueMu* GSS corresponds to a somatic insertion, with no phenotype.
   b. To determine if the GSS instead corresponds to a germinal insertion, purchase the corresponding 96-well *RescueMu* grid library plate (*see* **Subheading 2.8.**).

c. To screen the 96-well plate for a *RescueMu* germinal insertion, use a *RescueMu* read-out primer and a primer to the GSS of interest to screen by PCR using **steps 3–7** in **Subheading 3.8.3**. If the GSS of interest is found in both a row and column sample, then proceed to **step 6**.
d. If the GSS of interest is not found in both row and column samples, it is possible that a germinal insertion does exist, but was not retrieved during plasmid rescue in both row and column pools. To proceed, request *RescueMu* seed for all of the 48 plants in the row or column pool of the GSS. After growing these progeny, isolate leaf genomic DNA (**Subheading 3.2.**), then use PCR to screen leaves for the GSS-specific *RescueMu* insertion by following **Subheading 3.8.3., steps 3–7**. If an insertion is found, it is likely to be germinal, and thus, proceed to **step 7** of this section.

4. If multiple GSSs are retrieved, then click on the Sequence code of each GSS. At the bottom of each new page, note the Grid letter and Row/Column location.
5. If the multiple GSSs belong to only a row(s) or column(s) within a grid, but not both, then proceed with **step 5.** If the GSSs belong to both a row and column within a grid, then go to **step 6**.
   a. As the number of duplicate GSSs in only a row or column sample increases, the probability that the GSSs correspond to a germinal insertion increases.
   b. To determine if the GSS instead corresponds to a germinal insertion, purchase the corresponding 96-well *RescueMu* grid library plate (*see* **Subheading 2.8.**).
   c. To screen the 96-well plate for a *RescueMu* germinal insertion, use a *RescueMu* read-out primer and a primer to the GSS of interest to screen by PCR using **steps 3–7** in **Subheading 3.8.3**. If the GSS of interest is found in both a row and column sample, then proceed to **step 6**.
6. If the GSSs correspond to a row and column within a grid, then request seed for the *RescueMu* plant at the field grid intersection (**Fig. 2**).
7. Search for a visible phenotype in PhenotypeDB using Location Search by entering the Grid letter, Row, and Column numbers (*see* **Subheading 3.8.4.**).
8. Isolate genomic DNA from the candidate plant(s) and confirm the presence of a *RescueMu* germinal insertion by PCR using the appropriate *RescueMu* read-out primer and a gene-specific primer (*see* **Subheading 3.8.3.**, **steps 3–6**).
9. Perform a segregation analysis of the progeny by PCR to determine if the *RescueMu* allele cosegregates with the mutant phenotype.

3.8.6. How to Identify a RescueMu *Insertion Responsible for a MGDP Mutant Phenotype*

1. In the initial MGDP *RescueMu* grids, most mutants are caused by *MuDR/Mu* elements, not *RescueMu* (*see* **Note 2d**).
2. Use PhenotypeDB to locate the Grid letter, row and column numbers of the mutant (*see* **Subheading 3.8.4.**).
3. At the time of this protocol submission, the *RescueMu* GSS database could not be searched by row or column location. Instead, there is a link in PhenotypeDB

from any plant grid location to all GSSs found in the corresponding row and column. These may be queried against one another, but they may total hundreds of sequences. Consult PhenotypeDB or ZmDB for future upgrades.

4. Alternatively, request seed from the mutant plant (*see* **Subheading 2.8.**).
5. Isolate genomic DNA from a mutant plant from two leaves not likely to share a clonal sector and perform separate plasmid rescues (*see* **Subheadings 3.2.–3.6.**).
6. Perform restriction digests on the plasmids recovered from both leaves and electrophorese the plasmids on an ethidium bromide-stained gel. If any plasmids appear to be identical between the two leaves, these may be germinal *RescueMu* insertions and should be sequenced (*see* **Subheading 3.7.**).
7. To determine if a candidate *RescueMu* insertion is responsible for the mutant phenotype, perform a segregation analysis of the progeny by PCR to determine if the *RescueMu* allele co-segregates with the mutant phenotype (*see* **Subheading 3.8.3., steps 3–5**).

4. Notes

1. Because *RescueMu* seed is transgenic, permission is required for interstate shipment in the U.S. or to other countries from the Maize Coop Stock Center in Illinois, U.S.A. In the U.S.A., a letter of notification to ship or grow transgenic seed must be submitted to APHIS, which is a branch of the U.S. Department of Agriculture, 10–30 d in advance (http://www.aphis.usda.gov/ppq/biotech). A detailed *RescueMu* APHIS notification template letter can be obtained from the Maize COOP at (http://w3.aces.uiuc.edu/maize-coop/Aphis-notification.html). Combined with information found in ref. ***3***, these documents contain detailed information about the origin and construction of the *RescueMu* transgenes for permit documentation.
2. The utility of *RescueMu* is entirely determined by the choice of the starting plant material. The following parameters should be used when requesting or selecting *RescueMu* stocks:
   a. Of the original 20 *RescueMu* lines constructed, the most active lines are designated R3-4, R3-8, R3-13, and R3-17 ***(3).*** Line R3-8 has a complex transgene array and is a rich source of somatic transposition events, but it has a very low frequency of transmissable *RescueMu* insertions.
   b. From the original *RescueMu* lines, seed containing transposing *RescueMu* elements should be hand-selected based on the appearance of frequent single-cell red spotting on the kernel (aleurone) surface when viewed under a microscope ***(3)***; these represent *RescueMu* excisions from the anthocyanin *R* transcription factor. To observe such spots, seed lines should be in a *MuDR r C1* genetic background. *RescueMu* spotting is rarely observed on other tissues including leaves and anthers regardless of the anthocyanin genotype.
   c. *MuDR* is prone to epigenetic inactivation ***(1,2)***. Poorly spotted kernels or ears with few spotted kernels should not be chosen for plasmid rescue. Unfortunately, even spotted kernels may give rise to seedlings with inactive *RescueMu* elements. There is a tendency for kernels, leaf sheaths and leaves to turn solid

or patchy red when *MuDR*/*RescueMu* is in the process of epigenetic inactivation, though lack of a red color is not a reliable indicator of *RescueMu* activity. *Mu* epigenetic inactivation is lower when seed stocks are maintained by outcrossing; hence it is useful to have or request non-*Mutator r C1* seeds from the Maize Co-op Stock Center to serve as parents.

d. *RescueMu* seed lines may be in a background with 1–3 copies of *MuDR* (known as the Minimal Line) or 10–50 copies of *MuDR* (known as the Standard Line) (reviewed in refs. ***1,2***). In the Minimal Line background, *RescueMu* has excellent kernel spotting, high somatic transposition, low epigenetic inactivation, but extremely weak germinal transposition. In the Standard Line, *RescueMu* may have excellent kernel spotting, high somatic transposition, high rates of epigenetic inactivation, and a >10% frequency of *RescueMu* transposition. The advantage of the Minimal Line backgrounds is that they contain few *Mu1–Mu*8 elements, whereas Standard Lines may contain >100 *Mu1–Mu*8 elements. Hence, mutants found in Standard Lines have a low probability of being caused by a *RescueMu* element.

e. Because developmentally older leaves (higher up the plant) have a greater probability of epigenetic activation, it may be best to collect DNA from lower on the plant if somatic insertion events are of interest. It has been observed that the first 4 leaves are a rich source of *RescueMu* somatic transposition events.

f. Though *RescueMu* elements at their original and complex transgene integration site have a low frequency of germinal transposition, this frequency may increase to 20–100% after the *RescueMu* element has transposed to a new location ***(3)***. These seed stocks are designated tr-r*Mu* (transposed *RescueMu*), and are better stocks than the original *RescueMu* lines for germinal transposition, though they offer no advantages for late somatic transposition.

g. If tr-r*Mu* seed stocks are requested in which *RescueMu* has transposed from the original *RescueMu*/pAHC20 transgene array and the array has been outcrossed away, these seed lines should not be glufosinate herbicide-resistant. If using a tr-r*Mu* line, a molecular strategy must be designed to prevent repeated recovery of the tr-r*Mu* plasmid during plasmid-rescue experiments (*see* **Note 3**). The genomic DNA sequence flanking the tr-r*Mu* must be known in order to identify a unique restriction site to restrict recovery of this locus, or the DNA can be used as a hybridization probe on a colony-lift.

h. In coming years, new *RescueMu* seed lines will be available from the Maize Gene Discovery Project which have smaller *RescueMu* elements, will use kanamycin instead of ampicillin to reduce general laboratory plasmid contamination, will have better restriction sites for plasmid rescue, will use *Bronze2* instead of *R* as the excision marker, and will likely have fewer background *MuDR*/*Mu* elements. These lines will be called *MiniMu* and *MidiMu*. See (www.zmdb.iastate.edu) for updates and details.

3. Ampicillin-encoding plasmids rescued from *RescueMu* stocks can include the original *RescueMu* transgene array in addition to the genetically linked plasmid

pAHC20 encoding the herbicide-resistance gene *bar*, which is also ampicillin-encoding. The original *RescueMu* donor element may be recovered, because new germinal *RescueMu* insertions are caused by duplication rather than "cut and paste" transposition ***(1,2)***. Fortunately, there is a unique *Bgl*II site 100 bp downstream of *RescueMu* in the original *RescueMu* transgene and an additional site in pAHC20. Therefore, to bias against the recovery of these plasmids, after they ligate to form circular DNA, the DNA is restricted with *Bgl*II in order to linearize them and prevent their replication in *E. coli*. Some percentage of true *RescueMu* insertions containing flanking *Bgl*II sites will also be selected against using this method.

4. Because *RescueMu* plasmid rescue involves the use of highly competent *E. coli* cells with selection on ampicillin media, contaminating laboratory plasmids found in laboratory solutions or aerosols are a severe problem unless actively prevented. For example, if 1 μg of contaminating plasmid DNA was diluted into 10,000 L of water, then even 1 μL of this solution would result in a transformed *E. coli* colony. The following measures should be taken to prevent plasmid contamination:
   a. A separate plasmid-free bench area should be designated for the sole purpose of genomic DNA isolation and plasmid rescue. Other plasmid work should not be performed at the same time as this procedure.
   b. Fresh reagents (enzymes, buffers, alcohol, water) and solutions should be purchased, aliquoted for single-use experiments in sterile plastic tubes if possible, labeled as plasmid-free, and segregated from general laboratory use. Alternatively, water and solutions may be treated with activated charcoal and then filtered.
   c. Where possible, sterile plastic instead of laboratory glassware should be used. Mortars, pestles, spatulas, tube racks, and other reusable materials including pipetors should be treated with UV light for 1 min.
   d. Pipetors and the bench area may be treated with a 0.2% (v/v) HCl solution or DNA-Zap spray (Ambion, Austin, TX, USA). The bench coat should be changed frequently.
   e. Aerosol barrier pipet tips should be used throughout the procedure.
   f. To detect external plasmid contamination, two plasmid rescue controls should always be performed. First, a control omitting genomic DNA will indicate if one of the plasmid rescue solutions or the handling itself is creating plasmid contamination, while a control omitting ligase will indicate if the genomic DNA has been contaminated.
   g. To detect the extent of foreign plasmid contamination within a rescued DNA sample, random bacterial colonies should be restriction digested on an agarose gel. Because *RescueMu* is 4.7 kb, true rescued plasmids range in size from 5–25 kb. Small plasmids are an indication of external plasmid contamination. Alternatively, *RescueMu* plasmids may be positively identified using *RescueMu*-specific PCR amplication (*see* **Subheading 3.4.**).
5. Do not attempt to digest the genomic DNA with restriction digest for long peri-

ods of time or with excess enzyme (as when preparing a Southern blot), because the activity of contaminating nucleases will result in loss of the restriction fragment overhang and inhibit subsequent ligation.

6. I have found that genomic DNA purification using phenol–chloroform gives the highest transformation efficiency compared to sepharose and other minicolumn purification procedures.
7. For the ligation step, the concentration of genomic DNA is kept deliberately low (<20 μg/mL) to promote intramolecular ligation rather than intermolecular ligation (see ref. ***13***). Unfortunately, early *Arabidopsis* T-DNA rescue protocols called for much higher concentrations of genomic DNA.
8. Until now, only Electromax DH10B cells from Life Technologies have worked efficiently for plasmid rescue for the following reasons: *(i)* the electroporation efficiency of these cells is 10^{10} colonies/μg plasmid DNA; *(ii)* these cells carry the *mcrA, mcrBC, mrr,* and *hsdRMS* mutations, preventing methylated cytosine and adenine residues from plant genomic DNA from being restricted; *(iii)* the cells carry the *deoR* mutation, allowing them to accept plasmids as large as 150 kb; *(iv)* While DH5α cells also carry this mutation, although high efficiency transformation is limited to plasmids <30 kb in size; and *(v)* DH10B cells carry the *recA1* mutation, thus decreasing the frequency of recombination between the 215 bp terminal inverted repeats of *RescueMu*.
9. Carbenicillin (final 100 mg/L) should be substituted for ampicillin when preparing LB plates. Carbenicillin is more stable than ampicillin and will reduce satellite colony formation.
10. In addition to tissue containing inactive *RescueMu* elements, there are three major reasons for poor plasmid rescue efficiencies. First, the DNA pellet may be lost during the numerous genomic DNA precipitation steps. To help reduce loss of DNA, it may be useful to add 1 μL of glycogen as a carrier. Prior to transformation, an aliquot of the digested DNA should be measured to quantitatively assess the plasmid rescue efficiency. Second, the ATP in the ligase buffer may degrade due to numerous freeze–thaw cycles. The ligase buffer should be aliquoted into single-use tubes and frozen, and then thawed only once. Finally, even a slight loss in bacterial cell transformation competency will result in a significant decrease in the number of colonies recovered. For this procedure, cells should not be used if previously thawed, and the thawing and handling instructions of the manufacturer (Invitrogen or LIFE Technologies) should be followed exactly.
11. Up to 100–800 colonies are recovered on carbenicillin plates/μg of *Kpn*I-digested genomic DNA electroporated into 50 μL of Electromax DH10B cells, when the genomic DNA is isolated from an active *RescueMu* seedling. If the DNA is also digested with *Bgl*II to bias against recovery of plasmids from the *RescueMu* transgene array, the colony number drops to 20–300/μg of genomic DNA. Because there is no *Kpn*I site internal to *RescueMu*, these plasmids contain maize genomic DNA flanking both sides of *RescueMu* and are large plasmids (mean = 12 kb). If genomic DNA is only needed from one flank, an internal *Eco*RI site may be used for plasmid rescue instead of *Kpn*I, resulting in the recovery of smaller plasmids,

and therefore, a higher frequency may be expected (1500–4000 colonies/μg using *Eco*RI alone, and 150–1000 colonies/μg using both *Eco*RI and *Bgl*II). Alternatively, the MGDP uses a *Bam*H1/*Bgl*II double digest (**Fig. 1**) to permit digestion and selection to be performed in a single step; *RescueMu2* does not have an internal *Bam*H1 site, though *RescueMu3* does. There is a flanking *Bam*H1 site near *Bgl*II at *Lc::RescueMu*, which further helps to prevent recovery of the original *RescueMu* transgene array. Regardless of the enzyme combination chosen, it is important to note that the plasmids recovered after this step may still be from the original *RescueMu* transgene array, and further colony-lift hybridizations are recommended.

12. Colony-lift hybridizations are performed to screen for the recovery of true transposition events. In the first screen, colony hybridization to *RescueMu*2- and *RescueMu*3-specific probes specifically detect *RescueMu* plasmids and thus screen against recovery of foreign plasmid contamination and the ampicillin-encoding plasmid pAHC20 (the *bar* transgene). The purpose of the second screen is to avoid the recovery of *RescueMu* plasmids from the original *RescueMu::Lc*/pAHC20 transgene array. To do this, a set of filters is hybridized with the *RescueMu*-specific probes, while a duplicate set is hybridized to a combination of probes which flank *RescueMu* in the transgene array, namely the CaMV35S promoter and the maize *Lc*(*R*) cDNA. Colonies are then selected which hybridize to *RescueMu*, but do not hybridize to the CaMV35S/*R* probes. These colonies should represent new insertions of *RescueMu*.
13. Plasmids pRescueMu2,p RescueMu3, pMR15, and pMR17 and maps may be obtained from Virginia Walbot at Stanford University (walbot@stanford.edu) ***(3)***. The pR plasmid for the CaMV35S and maize *Lc(R)*probes may be obtained from Sue Wessler at the University of Georgia (sue@dogwood.botany.uga.edu) ***(11)***.
14. Recovered colonies may represent *RescueMu* somatic or germinal insertions. Typically, if 8–10 colonies are sampled by restriction digest on an agarose gel, unique plasmids represent late somatic insertions, whereas duplicated plasmids represent germinal or rare early somatic insertions ***(3)***. In the case of a germinal insertion, the same plasmid should be recovered from two separate leaves on the same plant.
15. Each GSS from the MGDP has a sequence I.D. code. Examples are 1006162C04.x2 1006 and 1008035A02.y1 1008. The 3 digits preceding the x or y digit or the period (e.g., C04 and A02) indicate the exact well location in the 96-well PCR library plate (e.g., C is a row and 04 is a column on the plate) (**Fig. 2**). The first 4 digits, sometimes repeated at the end (e.g. 1006 and 1008), also identify the field grid (e.g., G and I). X means that the read-out sequence was from the right *RescueMu* TIR (primer *Mu3*-R) while y refers to a read-out from the left TIR (primer *Mu3*-L) (*see* **Subheading 2.7.**). In addition, some sequence I.D.'s are longer, e.g., 1006162C04.1EL_x2 1006. The extra 3 digits (i.e., 1EL) indicate that the sequence from the recovered *RescueMu* plasmid came after one restriction site (*Bam*H1 or *Bgl*II) used during plasmid rescue, and hence this may not be a contiguous sequence. If the sequence came after a junction recognized as

the possible ligation between a *Bam*H1 overhang and a *Bgl*II overhang, then "2EL" is added to the GSS I.D. instead.

Acknowledgments

Special thanks to the dedicated members of the following MGDP laboratories from which further information may be obtained: Virginia Walbot (Stanford University), Sarah Hake and Michael Freeling (University of California, Berkeley), Laurie Smith and Robert Schmidt (University of California, San Diego), Vicki Chandler, David Galbraith, and Brian Larkins (University of Arizona, Tucson), Marty Sachs (University of Illinois, Urbana-Champagne), and Volker Brendel (Iowa State University).

References

1. Walbot, V. and Rudenko, G. N. (2002) *MuDR/Mu* elements of maize, in *Mobile DNA II* (Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A., eds.), American Society of Microbiology, Washington, D.C., Ch. 23.
2. Bennetzen, J. L., Springer, P. S., Cresse, A. D., and Hendrickx, M. (1993) Specificity and regulation of the *Mutator* transposable element system in maize. *Crit. Rev. Plant Sci.* **12,** 57–95.
3. Raizada, M. N., Nan, G.-L., and Walbot, V. (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. *Plant Cell* **13,** 1587–1608.
4. SanMiguel, P., Tikhonov, A., Jin, Y. K., et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768.
5. Christensen, A. H. and Quail, P. H. (1996) Ubiquitin promoter-based vectors for high level expression of selectable and/or screenable marker genes in monocotyledonous plants. *Transgenic Res.* **5,** 213–218.
6. Das, L. and Martienssen, R. (1995) Site-selected transposon mutagenesis at the *hcf106* locus in maize. *Plant Cell* **7,** 287–294.
7. Hanley, S., Edwards, D., Stevenson, D., et al. (2000) Identification of transposon-tagged genes by the random sequencing of *Mutator*-tagged DNA fragments from *Zea mays*. *Plant J.* **22,** 557–566.
8. Dellaporta, S. (1994) Plant DNA miniprep and microprep: versions 2.1–2.3, in *The Maize Handbook* (Freeling, M. and Walbot, V., eds.), Springer-Verlag, New York, pp. 522–525.
9. Saghai-Maroof, M. A., Soliman, K., Jorgensen, R. A., and Allard, R. A. (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* **81,** 8014–8018.
10. Franck, A., Guilley, H., Jonard, G., Richards, K., and Hirth, L. (1980) Nucleotide sequence of cauliflower mosaic virus DNA. *Cell* **21,** 285–294.
11. Ludwig, S. R., Habera, L. F., Dellaporta, S. L., and Wessler, S. R. (1989) *Lc*, a member of the maize *R* gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region. *Proc. Natl. Acad. Sci. USA* **86,** 7092–7096.

12. Grunstein, M. and Hogness, D. (1975) Colony hybridization: a method for the isolating of cloned DNA's that contain a specific gene. *Proc. Natl. Acad. Sci. USA* **72,** 3961–3965.
13. Rommens, C. M. T., Rudenko, G., Dijkwel, P. P., et al. (1992) Characterization of the *Ac/Ds* behaviour in transgenic tomato plants using plasmid rescue. *Plant Mol. Biol.* **20,** 61–70.

4

Precious Cells Contain Precious Information

Strategies and Pitfalls in Expression Analysis from a Few Cells

Isabelle M. Henry and Dina F. Mandoli

Summary

Expression analysis, often encompassed in the term "functional genomics," is the link between physiology and molecular biology. Often, specific physiological changes in plant development are due to a limited number of genes, expressed exclusively in very few cells of an organ or organism. Compounding the situation, these physiological changes may also be transient. Therefore, searching for the responsible genes, though exciting and necessary to understand important processes, is hindered primarily by the scarcity of "precious" cells in the desired physiological state. Used judiciously, molecular methods such as reverse transcription polymerase chain reaction (RT-PCR), microarray analysis, or subtractive hybridization allow analysis of rare or special cells. Each of these methods has its advantages and pitfalls. Working with precious cells entails special biological strategies to avoid excessive work in obtaining the data and misinterpretation of it. To illustrate the logic and methods involved in working with precious cells–tissues, we describe how subtractive hybridization followed by expressed sequence tag (EST) sequencing can be used to search for a few genes specific to a few available cells.

Key Words

EST, *Acetabularia*, developmental biology, mRNA extraction, suppressive subtractive hybridization, bioinformatics, normalization, RT-PCR

1. Introduction

The transcriptome is the dynamic link between the genome and the proteome ***(1)***. Now that the entire genome of several organisms has been fully sequenced, and their sequences are starting to be deciphered, expression analysis reemerges as an important link between molecular biology and physiology. Unraveling the networks of coordinated gene expression, which allow the genome to

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

respond dynamically to changes in the physiology and environment of an organism, permits a deeper understanding of how the transcriptome functions.

A challenge facing biologists is to take the expression of a myriad of individual genes and tease apart which regulate each particular pathway and how these pathways interact. Expressed sequence tags (ESTs) are fragments of mRNA that have been reverse transcribed into DNA and cloned. An EST library represents a pool of expressed genes from an organism, organ, or cell. They can be tailored to represent a specific time in development or a specific tissue. ESTs, thus, constitute an easy way to access information about the level of transcription of genes and the overall dynamics of gene transcription.

EST analysis is used either to randomly identify new proteins or to identify proteins expressed in particular cells or under particular circumstances. ESTs are especially well-suited to studying nonmodel organisms, because no genetic background is needed and virtually any sequence data is new. Indeed, many new proteins have been identified using this technique, some in plants. Successes include the identification of proteins involved in xylem formation in pine ***(2)*** and in poplar trees ***(3)*** or the identification of genes involved in different life cycle stages of the brown alga *Laminaria digitata* ***(4)***.

Some EST analyses consist of randomly sequencing all possible ESTs, usually with the goal of sampling all proteins in the organism (**Table 1**). This kind of analysis needs extensive equipment, such as automated DNA preparation and sequencing, as well as bioinformatics to concatenate the data. As expected, the more evolutionary distant the taxa are from land species, the fewer ESTs can be matched to known sequences using Basic Local Alignment Search Tool (BLAST) searches (**Table 1**). Of course, we can anticipate that matching frequency will increase as more sequences and species are added to the databases.

Arrays are a second powerful way to analyze nonsubtracted EST libraries. Unfortunately, such resources are not yet available for many plant species. Richmond and Somerville ***(5)*** have reviewed EST arrays of plants. Unless researchers are willing and financially able to create their own arrays, dealing with other species often necessitates a different approach.

Finally, selected EST libraries can be created to identify suites of genes, e.g., genes that are critical to a specific developmental process, involved in cellular responses to chemical or physical cues, or involved in the interactions with pathogens ***(6)***. Here, the goal is to identify a limited number of genes that are expressed in very defined conditions and that may interact with each other. For such a targeted analysis, fewer carefully chosen cells or parts of cells are required, and a smaller number of sequences will be analyzed in more detail. One way to isolate these transcripts involves the creation of subtracted cDNA libraries, which are enriched in stage or cell-type specific ESTs. With subtracted cDNA libraries, one can either randomly sequence ESTs or target just

Table 1
Statistics from Selected Plant EST Sequencing Projects

Organism-cell	No. of ESTs	% Unique[a]	% Hits[b]	Cut off[c]	Reference
Pine xylem	1097	80	55	Score > 80 P value < 0.01	***(2)***
Poplar	5692	47	63	Score > 100	***(3)***
Rice	29,000	44	Average 25	Score > 40 E value < 10 e-4	***(24)***
Laminaria sporophyte	493		48	Score > 200	
Laminaria gametophyte	412		39	Score > 200	***(4)***
Laminaria both		66			

[a]Percentage of nonredundant clones within the ESTs.
[b]Percentage of sequences that produced relevant hits in BLAST searches (as reported by the authors).
[c]See explanation of P and E value and score at (*http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head2*).

those clones with the expression pattern of interest for sequencing. Screening followed by directed sequencing is more efficient and has the added advantage of eliminating false positives.

Our laboratory has created two subtracted libraries of different ages of *Acetabularia acetabulum*. This giant unicellular marine green alga undergoes complex morphogenesis during development ***(7)***. At reproductive onset, it forms a unique apical structure or cap. Amputation experiments suggest that adult apices possess the transcripts needed for cap initiation, while juveniles do not (J. Messmer and D. F. Mandoli, unpublished). To isolate transcripts needed for cap initiation, we created two subtracted cDNA libraries, one enriched in adult transcripts, the other enriched in juvenile transcripts. Once the subtracted libraries were created, we randomly sequenced 1000 ESTs before confirming the differential expression of each EST. For *A. acetabulum*, this strategy is advantageous for several reasons. First, very little sequence data is available for this species, so almost any sequence is interesting and constitutes new data. For example, a few clones sequenced at random from an *A. acetabulum* cDNA library included a fragment of a nicotinamide dinucleotide transhydrogenase, which is an enzyme previously thought to exist only in animals ***(8)***. Second, this approach makes sense for organisms that are part of a branch of the Tree of Life for which little sequence information is available. Access to the complete genome of *Arabidopsis thaliana* and a multitude of sequences from other organisms, combined with more powerful bioinformatics

tools, has made the kind of information resulting from an EST analysis more meaningful. Finally, the decreased cost of sequencing, a consequence both of genome projects and advances in technology, makes this financially possible.

"Precious" cells are those that are difficult to obtain in large quantities. Often, these cells are also precious because of the information they contain. They can express genes specific to a developmental process, environmental or pathological response, or biochemical pathway. Therefore, it is often desirable or even necessary to work with precious cells, despite the limitations they entail (*see* **Note 1**). Restricted amounts of starting material also limit the number and size of the screens that are feasible. If the expression data needs confirmation, e.g., via Northern blot analysis, yield and quality of mRNA is important. It is not possible to approach the problem by doing many pilot experiments, because the cells are just too precious. It is often cost-effective to optimize the protocols, especially the mRNA extraction, using nonprecious cells before performing the final experiment on precious cells.

Our case clearly illustrates these problems. We performed a suppressive subtractive hybridization using adult *A. acetabulum*. Culture procedures had to be developed allowing for these unicells to grow axenically and synchronously ***(9)***. Careful documentation of the biology of this unicell was essential to know at what age they had to be harvested and from what portion of the unicell mRNA had to be extracted ***(10)*** (*see* **Note 2**). Although these unicells are huge (up to 3 cm long), ≥90% of the vol is occupied by a central vacuole resulting in a low yield of RNA per fresh weight of tissue. Previous work indicated that the mRNA for subtraction ideally should originate from the growing apices of juveniles and adults, making harvesting of the material very time-consuming and technically challenging. In the end, we compromised by not dissecting in the first round of analysis, but making mRNA from whole juveniles and adults.

Most importantly, the critical part of this kind of work has to be done before designing a molecular approach. One first has to acquire sufficient knowledge of the physiology and biology of the organism to best select the starting material, otherwise the molecular approach will become merely an expensive goose chase. Because none of the following techniques are unique to our analysis and most of them could become the objects of a separate protocol chapter, we focus here on the logic behind the steps involved in the analysis and how they can be applied to dealing with precious cells. The major molecular steps after biological optimization are mRNA extraction, cDNA synthesis, suppressive subtractive hybridization (SSH), cloning of the libraries, sequencing, and sequence analysis (**Fig. 1**).

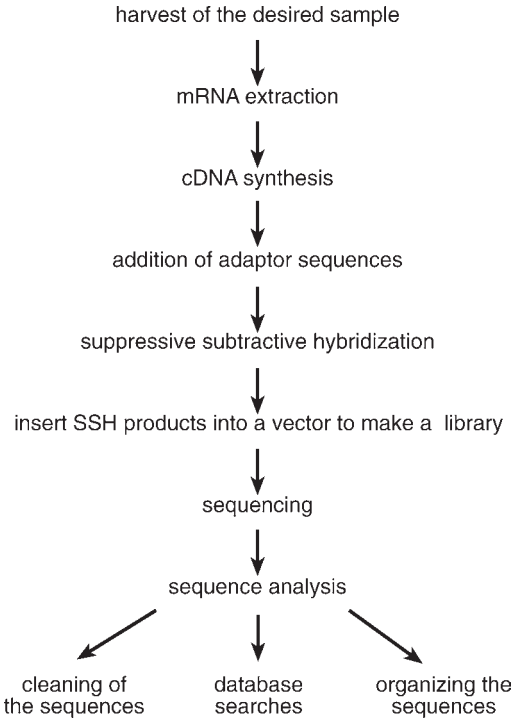


Fig. 1. Overview of the major steps leading to the EST analysis.

2. Material

2.1. mRNA

1. To avoid the action of RNAses, double-distilled sterile diethylpyrocarbonate (DEPC)-treated water (treated with 0.1% DEPC and autoclaved), sterile plasticware, and glassware baked at 150°C overnight should be used for the preparation of all solutions. All handling should be carried out wearing gloves.
2. Liquid nitrogen.
3. Mortar and pestle.
4. Extraction buffer: 2% hexadecyltrimethylammonium bromide (CTAB), 2% polyvinyl pyrrolidone (PVP) (K30), 100 m*M* Tris-HCl, pH 8.0, 25 m*M* ethylenediamine tetraacetic acid (EDTA), 2.0 *M* NaCl, 0.5 g/L spermidine. Dissolve in DEPC-treated water, mix, and autoclave. Add β-mercaptoethanol to 2% just before use.
5. Chloroform.
6. 10 *M* LiCl: made in DEPC-treated water and autoclaved.

2.2. cDNA Synthesis

Smart Polymerase Chain Reaction (PCR) cDNA Synthesis kit (Clontech Laboratories).

2.3. SSH

PCR-Select cDNA Subtraction kit (Clontech Laboratories).

2.4. Cloning of the Libraries

1. 100% and 80% Ethanol.
2. Phenol–chloroform (1:1).
3. 3 *M* Sodium acetate.
4. TE buffer: 10 m*M* Tris-HCl, 1 m*M* EDTA, pH 8.0.
5. 10× $MgCl_2$-free PCR buffer (Promega), *Taq* DNA polymerase 5 U/μL (Promega), 25 m*M* $MgCl_2$, and 10 m*M* dNTPs.
6. TOPO™-TA Cloning kit (Invitrogen).

2.5. Sequencing

1. Plasmid Miniprep Kit (Qiagen) for preparation of plasmid DNA.
2. Universal M13 primers for sequencing: M13F (5'-GTAAAACGACGGCCAG-3') and M13R (5'-CAGGAAACAGCTATGAC-3').
3. Model 3700 DNA Analyzer for separation and analysis of the sequencing reactions (Applied Biosystems).

2.6. Sequence Analysis

1. Sequencher™ (Genes Codes): for cleaning the sequences.
2. Perl: programming language.
3. Standard query language (SQL): to create our final database.
4. BLAST ***(11)***: to search the sequence databases.
5. Blastall: to search our own sequences.
6. InterPro: to search protein motif databases. InterPro is accessible on-line via the European Bioinformatics Institute at (http://www.ebi.ac.uk/interpro) ***(12)***.

3. Methods

The following section points out the major steps of this analysis, some of the problems inherent with each step, and presents resources available for troubleshooting. **Figures 2–5** give an overview of the different steps involved, from cDNA synthesis to cloning of the subtracted libraries. More detailed descriptions of each step can be found in each kit's user manual.

3.1. mRNA Extraction

Plant and parts of plants differ widely in their polysaccharide and polyphenolic content, making it necessary to adjust the extraction protocols almost on

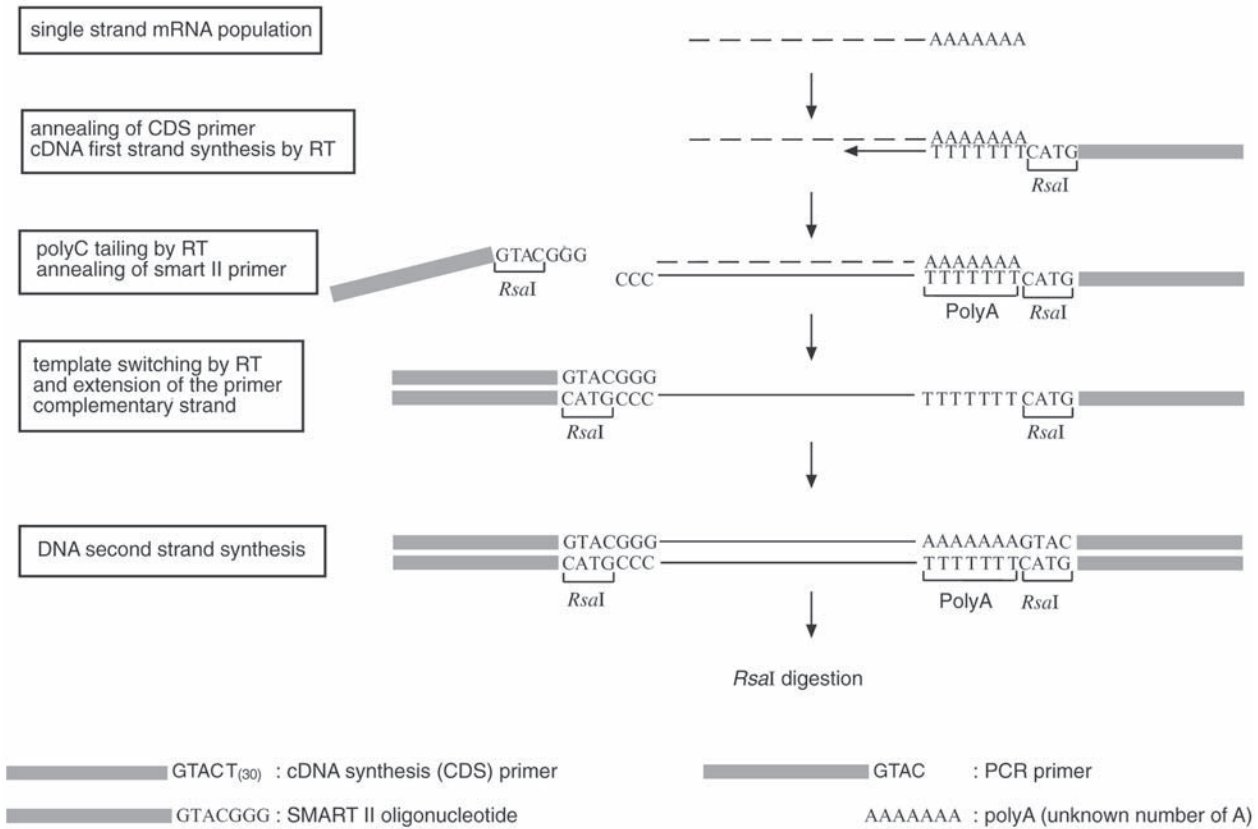


Fig. 2. Overview of the steps involved in cDNA synthesis starting from total RNA or mRNA (adapted from Smart PCR cDNA Synthesis kit user manual).

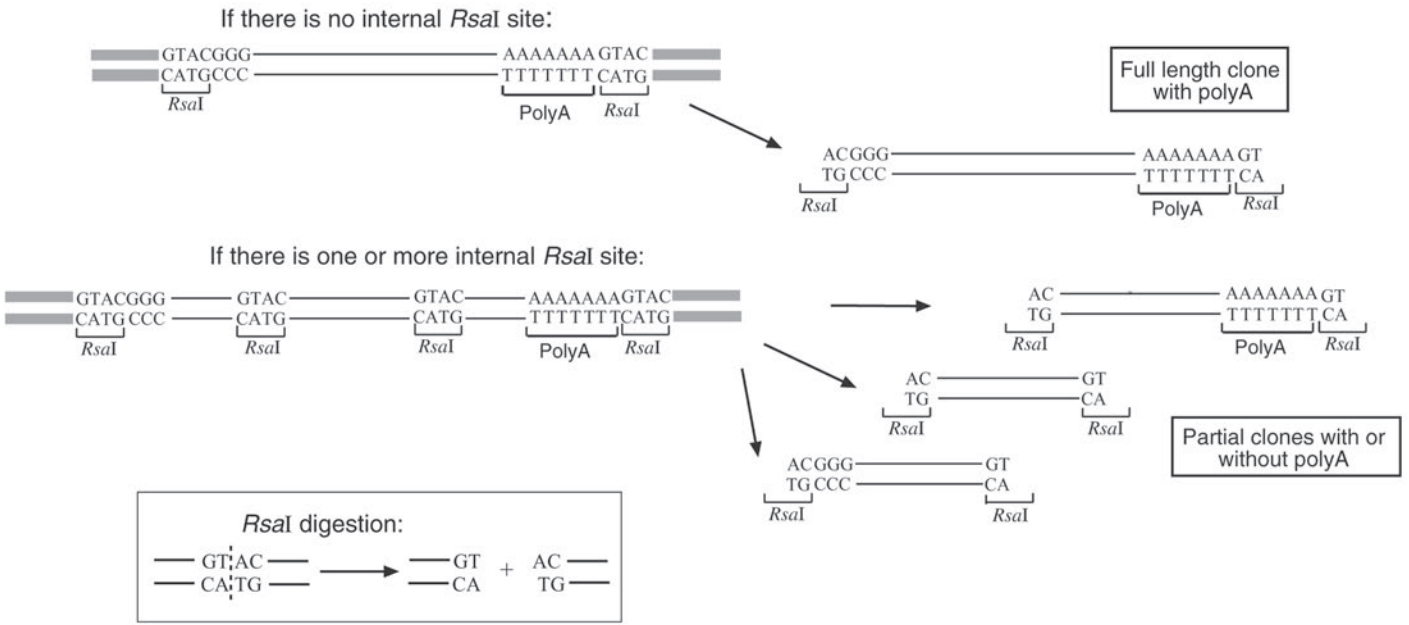


Fig. 3. *Rsa*I restriction of the double-stranded cDNA population.

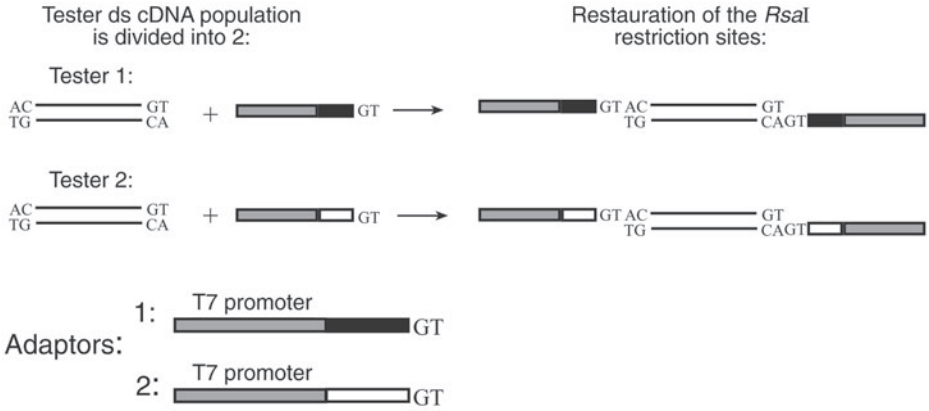


Fig. 4. Addition of adaptor sequences to the *Rsa*I restriction fragments.

a case-to-case basis (*see* **Note 3**). Below is the protocol we used to extract RNA from frozen *Acetabularia* cells. References to other protocols are listed in **Note 4**. This protocol is based on Chang et al. ***(13)***:

1. Grow *Acetabularia* synchronously in artificial seawater until they reach the desired developmental age. Axenic cultures are obtained by decontaminating the mature caps and using the resulting gametangia for mating. The zygotes produced are then grown in our sterile artificial seawater (Ace-27 ***[14]***), under cool white fluorescent lights at a photon flux density of 170 μmol/m^2s on a 14-h light/10-h dark photoperiod at 21°C. *Acetabularia* cells are repeatedly diluted according to their developmental age ***(14)***.
2. Harvest by filtration or using sterile dental tools, dry briefly on a Kimwipe, weigh on aluminum foil, and freeze each packet of unicells in liquid nitrogen.
3. Grind 5–10 g (*see* **Note 5**) of frozen material under liquid nitrogen to a fine powder using a chilled mortar and a pestle.
4. Transfer the ground cells to an Oakridge tube containing extraction buffer (dissolve 0.1–0.2 g of ground cells/mL of extraction buffer) heated to 75°C. Incubate at 75°C for 5–30 min.
5. Add 1 vol of chloroform, heated to 75°C, and mix well by shaking.
6. Centrifuge for 15 min at 22°C and 12,000*g*.
7. Transfer the top phase to a fresh tube.
8. Add 1 vol of chloroform, heated to 75°C, mix well by shaking.
9. Centrifuge for 15 min, at 22°C and 12,000*g*.
10. Transfer the top phase to a fresh tube.
11. Add one-fourth vol of 10 *M* LiCl.
12. Pack the tubes in ice and place the ice bucket at 4°C overnight.
13. Centrifuge for 20 min, at 4°C and 17,400*g*.
14. Discard the supernatant and dry the RNA on the bench.

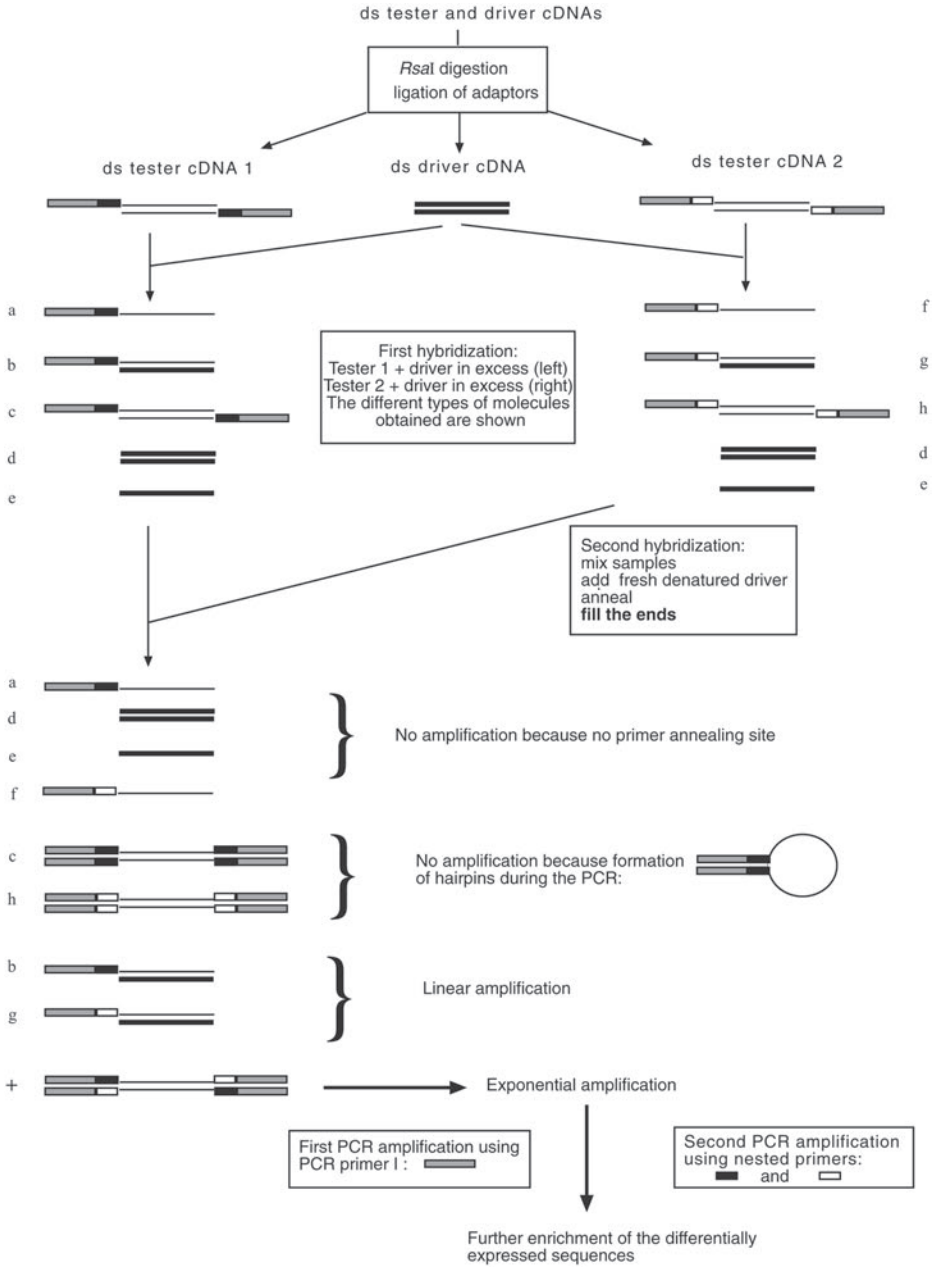


Fig. 5. Major steps in SSH (adapted from PCR-Select cDNA Subtraction kit user manual).

15. Add 500 µL of DEPC-treated water and heat the sample to 65°C for 20 min.
16. Transfer the sample to a 1.7 m*L* microfuge tube.
17. Make 1:50 and 1:100 dilutions and measure A_{260} and A_{280} to determine the quality and quantity of the RNA (*see* **Note 6**).
18. Store the samples at –80°C.

3.2. cDNA Synthesis

The detailed protocol can be found in the Smart PCR cDNA Synthesis kit user manual. **Figure 2** summarizes the steps involved in cDNA synthesis:

1. cDNA first strand synthesis by reverse transcription (RT), using the CDS primer (5'-AAGCAGTGGTAACAACGCAGAGTACT$_{(30)}$N$_{-1}$N-3').
2. polyC tailing by RT, annealing of the smart II primer (5'-AAGCAGTGGTA ACAACGCAGAGTACGCGGG-3'), followed by extension of the primer complimentary strand after RT has switched templates.
3. Synthesis of the second cDNA strand.

The primers used for cDNA synthesis (CDS) and smart II oligonucleotide primers both contain an *Rsa*I restriction site. This results in a high percentage of restriction fragments containing a poly-adenylated [poly(A)] tract. Some clones of the final library, subtracted or not, are thus also likely to contain a poly(A) tract.

3.3. SSH

Subtractive hybridization allows one to "subtract" two mRNA populations, i.e., to find genes that are expressed exclusively in one mRNA population (the tester population) and not in the other (the driver population). Therefore, the choice of the driver and tester mRNA populations is crucial (*see* **Note 7**).

The detailed protocol can be found in the PCR-Select cDNA Subtraction kit user manual. The following steps are summarized in **Figs. 3–5**:

1. *Rsa*I digestion: the populations of cDNAs are cut by a restriction enzyme (**Fig. 3**). In our case, the enzyme is *Rsa*I (*see* **Note 8**).
2. Creation of the final tester populations by dividing the tester *Rsa*I fragment population into two pools and adding a different set of adaptors (**Fig. 4**) to each pool. The driver population is not modified. This is the innovative step that makes the subsequent PCR suppressive ***(15)***.
3. Subtractive hybridization: this involves several steps (**Fig. 5**).
   a. The two tester populations are each hybridized with an excess of driver population.
   b. The two pools are then mixed together, again with an excess of driver cDNA. Only fragments that remained single stranded in both pools will form duplexes bordered by two different adaptors.

c. The resulting population is then selectively amplified using primers that anneal to these adaptors. Consequently, only sequences that are bordered by two different adaptors will be amplified exponentially. The other fragments will not be amplified or will only be amplified in a linear fashion for different reasons (*see* **Fig. 5**).

Unfortunately, SSH can generate false positives (*see* **Note 9**). Different methods allow verification of the expression patterns of the candidate clones (*see* **Note 10**).

3.4. Cloning of the Libraries

1. Precipitate the library DNA by adding one-tenth vol of 3 *M* sodium acetate and 2 vol of 100% ethanol. Centrifuge at maximum speed for 5 min. Discard the supernatant. Rinse the pellet with 80% ethanol. Air-dry the pellet.
2. Resuspend the DNA in 25 µL of PCR cocktail: 2.5 µL of 10× buffer, 1.5 µL $MgCl_2$, 2 µL 10 m*M* dNTPs, 18.875 µL water, and 0.125 µL *Taq* DNA polymerase. Incubate at 72°C for 8–10 min. This will add 3' A-overhangs to the PCR products for subsequent cloning into the TOPO TA-cloning vector.
3. Extract immediately with an equal vol of phenol–chlorofom: add one-tenth vol of 3 *M* sodium acetate and 2 vol of 100% ethanol. Precipitate the DNA by centrifuging for 5 min at maximum speed in a microcentrifuge. Discard the supernatant, rinse the pellet with 80% ethanol, and air-dry the pellet.
4. Resuspend the DNA in TE buffer to the starting vol of the DNA amplification reaction.
5. The DNA product is now ready to be cloned into the TOPO TA-cloning vector, following the manufacturer's recommendations.

3.5. Sequencing

1. Randomly pick 1000 clones using sterile toothpicks (500 from the adult-enriched library and 500 from the juvenile-enriched library).
2. Grow the bacteria overnight in 5 mL of LB medium at 37°C.
3. Prepare the plasmid DNA using Plasmid Miniprep kits. Resuspend the plasmid DNA into 2× 40 µL EB buffer (10 m*M* Tris-Cl, pH 8.5).
4. The sequencing reactions are as follows: mix 400 ng of plasmid DNA and 4 pmol of primer (M13R or M13F) in 10 µL of water. Add 2 µL BigDye™ Terminator mixture, version 2 (Applied Biosystems), 4 µL BetterBuffer (The Gel Company), and 4 µL water.
5. The cycling parameters are those recommended by BigDye Terminator, except that the reactions run for 35 cycles instead of 25.
6. Reactions are cleaned up with the Multiscreen/Sephadex® Procedure (Millipore) according to the manufacturers instructions (Millipore Technical Note TN053).
7. The resulting 20 µL of cleaned up sequencing reaction product (in water) is placed on a Model 3700 DNA Analyzer for separation and analysis (*see* **Note 11**).

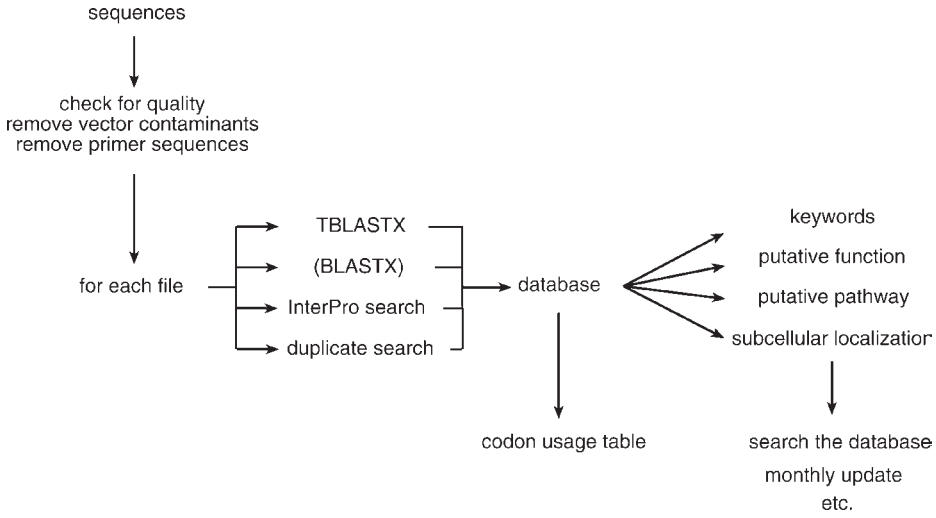


Fig. 6. Outline of the steps involved in the analysis of EST sequences.

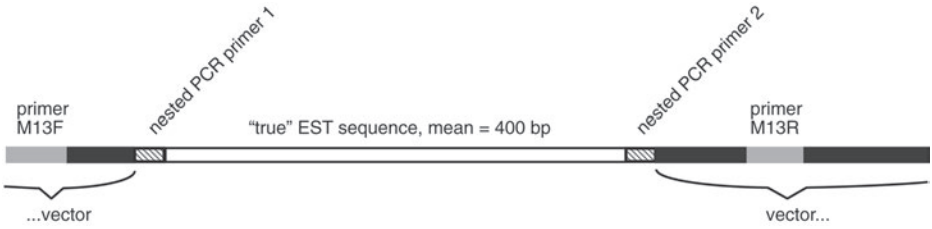


Fig. 7. Structure of a typical EST sequence before the cleaning steps.

3.6. Sequence Analysis

The following paragraphs describe our sequence analysis (**Fig. 6**) and the limited interpretation that it can provide.

3.6.1. Cleaning of the Sequences

A typical sequence is presented in **Fig. 7**. Only the true EST sequence should remain after the electronic "cleaning" of the sequences. This is critical to avoid erroneous BLAST results (*see* **Note 12**).

1. Removal of the vector: align each new sequence with that of the two ends of the vector and trim any part of the new sequence that is vector sequence. In general, cleaning sequences one by one is preferable for small data sets, because it allows more reliable removal of contaminant flanking sequences and visualization of

the sequences. Alternatively, this can be done using sequence analysis software such as Sequencher. If vector sequence is found at both ends of the sequences, the full clone has been sequenced. If not, one might want to sequence the clone from the other end, depending on how much usable sequence has been obtained from the first sequence run.

2. Removal of the primer sequences: The primer sequences from SSH amplifications should also be removed. In our case, these are nested PCR primer 1 (5'-TCGAGCGGCCGCCCGGGCAGGT-3') and nested PCR primer 2 (5'-AGCGT GGTCGCGGCCGAGGT-3').
3. Handling of the sequences: once the sequences are clean, they can be pasted into a single file for subsequent batch analysis. Using a Fasta format (*see* **Note 13**) is preferable, because most bioinformatics software accepts this format as input.

3.6.2. Database Search

It is best to use programs that allow analysis of batches of sequences.

1. TBLASTX searches: this search translates each input sequence into all six possible reading frames and compares the resulting protein sequences against the nucleotide database, which is also translated into all six possible reading frames. This kind of analysis is computationally intensive, but is the most likely to generate a maximum number of hits (*see* **Note 14**). The parameters we used in the TBLASTX searches were:
   a. Database to search: nt ("nonredundant" nucleotide sequences).
   b. Number of descriptions: 50.
   c. Number of alignments: 10.
   d. Expect value: 0.0001.
   e. The rest of the parameters were kept on the default setting.
2. BLASTX searches: knowing what genetic code the organism of interest uses is essential for sequence analysis, because TBLASTX searches do not allow you to specify the genetic code (*see* **Note 15**). For those organisms, performing a BLASTX search is also necessary. The parameters we used in the BLASTX searches were:
   a. Database to search: nr ("nonredundant" protein sequences).
   b. Number of descriptions: 50.
   c. Number of alignments: 10.
   d. Expect value: 0.0001.
   e. Code: 6 (corresponds to the ciliate genetic code).
   f. The rest of the parameters were kept as defaults.
3. InterPro searches: the remaining clones can be compared to protein signature databases. There are several such databases, but most of them have recently been merged into InterPro. This database was developed to create a single coherent resource for diagnosis and documentation of protein families. So far, it contains data from the Pfam database (divergent domains), PROSITE (functional sites), PRINTS (protein families), ProDom (cluster database, derived automatically

Table 2
Two Organizations of ESTs or Genes into Functional Classes

Convention established by TIGR in the expressed gene anatomy database (EGAD) ***(25)***	The "12 functional groups" based on catalogues established for *Escherichia coli, Saccharomyces cerevisiae,* and *A. thaliana* ***(3)***
Cell division.	Cell cycle control.
Cell signaling and communication.	Signal transduction.
Cell structure and mobility.	Cytoskeleton.
	Cell wall formation.
Cell or organism defense.	Stress-related proteins.
Gene or protein expression.	Proteins synthesis.
	Proteins modification, degradation, or targeting.
	DNA binding proteins.
Ribosomal proteins.	Nucleotide and amino-acids metabolism.
Metabolism (including photosynthesis).	Metabolism.
	Hormone synthesis-related.
	Other proteins.
	Unknown (similar to uncharacterized sequences).
	No hits.

from sequence databases), and BLOCKS (ungapped multiple alignment of protein families).

4. Update of the results: if the purpose of this analysis is to obtain a database of sequences from a given organism, it is useful to keep blasting the "no hit" clones every month against the "month" database (which is the database containing only the newly released sequences) to search for new entries.

3.6.3. Organizing the Sequences

1. Identification of the duplicates: sequences can be organized into contigs or nonredundant groups by placing duplicate or overlapping sequences together. The number of clones that fall into no contig, so called "singletons," is always overestimated, because some will be nonoverlapping sequences of the same transcript. Several software packages are available for contig analysis: Sequencher; Blastall, using a personal database ***(11)***; or The Institute for Genomic Research (TIGR) assembler ***(16)***. Using the results of the BLAST searches, ESTs can then be classified into functional groups (**Table 2**).

2. Creation of an EST database: this database will contain all the information collected about each EST, along with its sequence, keywords signifying putative function, pathway, subcellular localization, etc. (**Fig. 6**). It can range from a simple Excel® file, to a database that can be searched for these keywords and can be automatically updated using BLAST searches on a regular basis.

4. Notes

1. Dealing with precious cells is often complicated by various factors such as:
   a. Contaminants: contamination results either from other organisms, from other tissues in the same organism, or from the same type of cell at a different stage in the life cycle. Therefore, it is critical that organisms–cells are as clean and as developmentally synchronous as possible.
   b. Time: harvesting precious cells can be extremely time-consuming. Unfortunately, once the cells have reached the desired developmental age, it is often difficult to be able to harvest them quickly enough. Therefore, this should be a consideration in designing the overall molecular approach.
2. The results of the SSH can vary greatly with a little change in the choice of the starting material. It is, therefore, necessary to focus on the physiology and biology of the system before selecting both the experimental and control starting material for molecular analysis. Such factors include: circadian time, i.e., time of day relative to the light–dark cycle, photon flux density, spectral environment, e.g., UV component, temperature during day and night, water status, nutrient status, season of the year, developmental age of the organism, portion or subcellular portion of the organism, population density, time postexperimental treatment, and ecological considerations, such as microbes, pathogens, etc.
3. Good quality RNA is critical for the success of all subsequent steps in making and for expression analysis of the ESTs. Unfortunately, plant cells are surrounded by a cell wall composed primarily of carbohydrates that tend to co-precipitate with RNA. Similarly, phenolic compounds in plants also tend to co-precipitate with RNA. Most plant cells also contain large aqueous vacuoles, which lower the yield of RNA extracted/fresh weight. Only when we were able to consistently obtain a good yield of quality mRNA did we apply this method to our precious cells and proceed to the next step. We advocate trial runs on tissue that is as similar as possible to the target tissue.
4. Speirs and Longhurst ***(17)*** have compared RNA extraction protocols and methods, including approximate yields of RNA. They also list the tissue(s) used with these methods. More recent publications present the following protocols or techniques: *(i)* homogenization methods ***(18)***; *(ii)* benzyl chloride extraction from rice leaves of different ages ***(19)***; *(iii)* extraction from succulent plant species ***(20)***; and *(iv)* extraction from plants containing high levels of phenolics or polysaccharides ***(21)***. Finally, different RNA extraction protocols are listed at (http://www.protocol-online.org/prot/Molecular_Biology/RNA/RNA_extraction/index.html).
5. We used 7.15 g of juvenile cells (approx 18,000 cells) and 9.15 g of adult and reproductive cells (approx 2000 cells).

6. We typically obtain 20–30 μg of total RNA from 1 g of adults. This number is probably higher for juveniles. We only extracted RNA from juvenile cells once obtaining 216 μg of total RNA from 1 g fresh weight.
7. Variability in gene expression: if the library has been constructed from cells of only one organism, the natural variation in the level of expression of numerous genes between different individuals might be misleading and result in a high percentage of false positives in a SSH library. Mice that are genetically identical and have almost identical environmental histories show surprising variability in levels of gene expression. Disturbingly, several genes, which varied "normally" in this study, had been previously interpreted as differentially expressed in experiments comparing cells or tissues exposed to variable conditions *(22)*. Ideally, consensus will be reached as to which genes within a certain organism are subject to such variation, and databases of those genes will result allowing researchers searching for differentially expressed genes to focus on differences particular to their experimental manipulations (D. A. Coil, personal communication). To the best of our knowledge, no such analysis has been performed on a plant species so far, but it would be surprising if this phenomenon were limited to animals.
8. The restricted fragments can include full-length clones or not, according to whether the initial cDNA possesses an internal *Rsa*I site or not (**Fig. 3**).
9. For various reasons, SSH is never 100% successful, i.e., it generates a varying percentage of true positives. According to the PCR-Select cDNA Subtraction kit user manual, it can be as high as 95% or as low as 5%. Application of different methods allowing true quantitative expression analysis are, therefore, crucial.
10. Performing a subtractive hybridization decreases the number of clones to be sequenced, but does not provide definitive data as to the expression pattern of a particular gene. Final determination of the "true positives" requires one or more of the following methods: Northern blots, quantitative or competitive PCR, or quantitative ribonuclease assays. Quantitative PCR is probably a method of choice when dealing with precious cells, because it only requires small quantities of starting material.
11. Sequence quality: usually and for economical reasons, ESTs are only sequenced once. Consequently, the sequence data is never 100% accurate and decreases the quality of the public databases. Resequencing genes of interest is essential to further work.
12. Most databases (protein or nucleotide) are noncurated, i.e., they rely solely on the expertise of the people using it and adding to it. A simple homology search shows that many GenBank® entries are not devoid of vector sequence. For example, if the polylinker site of our cloning vector is used in a BLASTN analysis, the first 100 hits have E values of 2e –31 or less, and 39 of them are not described as vector sequence. Hence, the presence of vector sequence in the EST generates many erroneous BLAST results. It is prudent to know the limitations of the tools one is using, and this applies to bioinformatics tools used for homology searches: a simple query will indicate if and how a database you are using is curated.
13. Fasta format: a sequence in fasta format begins with a single-line description of

the sequence, followed by lines of sequence data. The description line is characterized by a greater than (>) symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. More than one sequence can be included in the same file. An example sequence in fasta format is:
> name of sequence number 1
ATGCATGAGCTCGATCGAGTCGATTAGCTAGCTAGGACTCA
GCTACGACTACGACTACAGCGACTACG
> name of sequence number 2
ATGATGATTGATTAGATAACGCTGCATTACGCATCAGCATCTCAGCT
ACAGCACTACACATCAGCAGCTCA

14. Depending on the library and the organism it was derived from, the number of ESTs producing relevant BLAST hits varies widely (**Table 1**). For organisms such as *Acetabularia*, for which no closely related taxa have been extensively sequenced, the percentage of ESTs producing relevant hits is usually <50%. Indeed, only approx 20% of the *Acetabularia* ESTs generated relevant hits when used in BLAST searches. They remain "novel" for the time being.
15. *Acetabularia **(23)***, like the ciliates, does not use the universal genetic code but uses a different one of the 17 genetic codes known in the Tree of Life (http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c#SG2). For such organisms, it is necessary to also perform a BLASTX analysis. This search translates the input sequences into all six possible reading frames and compares the resulting protein sequences against a nonredundant protein database. This search allows you to choose which genetic code should be used for the translation of the query. Oddly, TBLASTX also allows you to specify the genetic code but that information is not used in the analysis.

Acknowledgments

We thank Dr. Szusanna Schwartz-Sommer (Max-Planck-Institute für Zuchtungsforschung, Cologne, Germany) for constructing the subtracted cDNA pools and for all her advice. We thank Mark Wilkinson (Founder, Illuminae Media) for help with the bioinformatics tools, the format of the final database, and useful comments. We thank Erich Grotewold, Marcela Hernandez, and the Plant-Microbe Genomics Facility (The Ohio State University, Columbus, OH) for performing the sequencing for our EST project.

References

1. Nelson, P., Han, D., Rochon, Y., et al. (2000) Comprehensive analyses of prostate gene expression: convergence of expressed sequence tag databases, transcript profiling and proteomics. *Electrophoresis* **21,** 1823–1831.
2. Allona, I., Quinn, M., Shoop, E., et al. (1998) Analysis of xylem formation in pine by cDNA sequencing. *Proc. Natl. Acad. Sci. USA* **95,** 9693–9698.
3. Sterky, F., Regan, S., Karlsson, J., et al. (1998) Gene discovery in the wood-

forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **95,** 13330–13335.
4. Crepineau, F., Roscoe, T., Kaas, R., Kloareg, B., and Boyen, C. (2000) Characterization of complementary DNAs from the expressed sequence tag analysis of life cycle stages of *Laminaria digitata* (Phaeophyceae). *Plant Mol. Biol.* **43,** 503–513.
5. Richmond, T. and Somerville, S. (2000) Chasing the dream: plant EST microarrays. *Curr. Opin. Plant Biol.* **3,** 108–116.
6. Carulli, J., Artinger, M., Swain, P. M., et al. (1998) High throughput analysis of differential gene expression. *J. Cell Biochem.* **30/31,** 286–296.
7. Mandoli, D. (1998) Elaboration of body plan and phase change during development of *Acetabularia*: how is the complex architecture of a giant unicell built? *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49,** 173–198.
8. Arkblad, E., Betsholtz, C., Mandoli, D., and Rydstrom, J. (2001) Characterization of a nicotinamide nucleotide transhydrogenase gene from the green algae *Acetabularia acetabulum* and comparison of its structure with those of the corresponding genes in mouse and *Caenorhabditis elegans*. *Biochim. Biophys. Acta* **1520,** 115–123.
9. Hunt, B. and Mandoli, D. (1996) A new artificial seawater that facilitates growth of large numbers of cells of *Acetabularia acetabulum* (Chlorophyta) and reduces the labor inherent in cell culture. *J. Phycol.* **32,** 483–495.
10. Runft, L. and Mandoli, D. (1996) Coordination of cellular events that precede reproductive onset in *Acetabularia acetabulum*: evidence for a "loop" in development. *Development* **122,** 1187–1194.
11. Altschul, S., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
12. Apweiler, R., Attwood, T. K., Bairoch, A., et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40.
13. Chang, S., Puryear, J., and Cairney, J. (1993) A simple efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11,** 113–116.
14. Mandoli, D. (1998) What ever happened to *Acetabularia*? Bringing a once-classic model system into the age of molecular genetics. *Int. Rev. Cytol.* **182,** 1–67.
15. Diatchenko, L., Lau, Y. F., Campbell, A. P., et al. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. USA* **93,** 6025–6030.
16. Sutton, G., White, O., Adams, M., and Kerlavage, A. (1995) A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1**, 1–19.
17. Speirs, J. and Longhurst, T. (1993) RNA extraction and fractionation, in *Methods in Plant Biochemistry, Vol. 10*, (Bryant, J., ed.), Academic Press, San Diego, pp. 1–32.
18. Eggermont, K., Goderis, I., and Broekaert, W. (1996) High-throughput RNA extraction from plant samples based on homogenisation by reciprocal shaking in the presence of a mixture of sand and glass beads. *Plant Mol. Biol. Rep.* **14,** 273–279.

19. Suzuki, Y., Makino, A., and Tadahiko, M. (2001) An efficient method for extraction of RNA from rice leaves at different ages using benzyl chloride. *J. Exp. Bot.* **52,** 1575–1579.
20. Gehrig, H., Winter, K., Cushman, J., Borland, A., and Taybi, T. (2000) An improved RNA isolation method for succulent plant species rich in polyphenols and polysaccharides. *Plant Mol. Biol. Rep.* **18,** 369–376.
21. Salzman, R., Fujita, T., Zhu-Salzman, K., Hasegawa, P. M., and Bressan, R. A. (1999) An improved RNA isolation method for plant tissues containing high levels of phenolic compounds or carbohydrates. *Plant Mol. Biol. Rep.* **17,** 11–17.
22. Pritchard, C., Hsu, L., Delrow, J., and Nelson, P. S. (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci. USA* **98,** 13266–13271.
23. Schneider, S., Leible, M., and Yang, X. (1989) Strong homology between the small subunit of ribulose-1,5-biphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. *Mol. Gen. Genet.* **218,** 445–452.
24. Yamamoto, K. and Sasaki, T. (1997) Large-scale EST sequencing in rice. *Plant Mol. Biol.* **35,** 135–144.
25. Adams, M., Kerlavage, A. R., Fleischmann, R. D., et al. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377,** 3–174.

5

Combined ESTs from Plant–Microbe Interactions

Using GC Counting to Determine the Species of Origin

Edgar Huitema, Trudy A. Torto, Allison Styer, and Sophien Kamoun

Summary

A diversity of microorganisms establishes intimate associations with plants. Whether pathogenic or symbiotic, such interactions are the result of complex recognition events between plants and microbes, leading to signaling cascades and regulation of countless genes involved in the interaction. A key step in unraveling the mysteries of plant–microbe interactions lies in defining the transcriptional changes that occur in both the host and the microbe during their association. The sum of the transcripts, from both host and microbe, which are produced during their association, has been defined as the interaction transcriptome. One approach to analyze interaction transcriptomes is to perform large-scale sequencing of cDNAs (expressed sequence tags or ESTs) obtained from infected plant tissue and representing a mixture of host and microbe sequences. In some cases, the two organisms have markedly different GC content, allowing most ESTs to be easily distinguished on this basis. In this chapter, we describe a GC counting method to determine the species of origin of ESTs obtained from interactions between plants and oomycetes or other high GC content microbes.

Key Words

plant–microbe interactions, *Phytophthora*, oomycetes, interaction transcriptome, EST annotation, GC content, GC counting

1. Introduction

A diversity of microorganisms establishes intimate associations with plants. Whether pathogenic or symbiotic, such interactions are the result of complex recognition events between plants and microbes, leading to signaling cascades and regulation of countless genes involved in the interaction. A key step in

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

unraveling the mysteries of plant–microbe interactions lies in defining the genetic components involved and the transcriptional changes that are occurring in both the host and the microbe ***(1)***. The sum of the transcripts, from both host and pathogen, which are produced during their association, has been defined as the "interaction transcriptome" ***(1)***. Each interaction transcriptome has been hypothesized as being unique to a particular host–pathogen or host–symbiont association, and its characterization should help to define the complex mechanisms involved in establishing and maintaining their interaction ***(1)***.

The emergence of low-cost high-throughput DNA sequencing methods has allowed plant biology to enter the era of genomics. In particular, projects involving large-scale sequencing of cDNAs (expressed sequence tags or ESTs) are ongoing for a wide variety of plants and plant-associated microbes. Similarly, ESTs generated from mRNA isolated from plant tissue infected with microbial pathogens have emerged as useful data sets for dissecting interaction transcriptomes ***(1,2)***. For example, this approach has been used for two eukaryotic microbial pathogens, the oomycete *Phytophthora infestans*, which causes late blight on tomato and potato (E. Huitema and S. Kamoun, unpublished; B. Baker et al. NSF Potato Genomics Project, www.tigr.org/tdb/potato), and *Phytophthora sojae*, which causes root and stem rot on soybean ***(1,2)***. ESTs generated from cDNA libraries constructed from *Phytophthora*-infected plant tissue could be of either pathogen or host origin. Thus, the challenge is to distinguish between the plant and *Phytophthora* EST populations using sequence analyses. In this case, plant and *Phytophthora* ESTs have markedly different GC contents, allowing most ESTs to be easily distinguished on this basis. For example, the percentage of GC content was assessed for sequences from cDNA libraries derived solely from *P. sojae* and soybean ***(2)***. Both sets of sequences produced distinct slightly overlapping normal distribution curves, with the pathogen ESTs averaging 58% GC content, and the host ESTs averaging 46% GC content ***(2)***. A similar analysis of sequences from a *P. sojae*-infected soybean cDNA library revealed ESTs to be clustered around two peaks corresponding to 46 and 58% GC content, suggesting that about two-thirds of the ESTs from this library are likely to be from the pathogen ***(2)***. In this chapter, we provide step-by-step instructions on how to run the GC counting method to help distinguish between host and microbe sequences from ESTs from interactions between plants and oomycetes or other high GC content microbes.

2. Materials

2.1. Hardware and Operating System

A workstation running the Linux operating system. For example, we currently use a Pentium III personal computer (PC) running Red Hat Linux OS.

2.2. Software

The GC counting program *GC* can be downloaded from (http://www.oardc.ohio-state.edu/phytophthora/gc.htm). The program was written in C++ and was only tested on the Linux platform.

Microsoft® Excel® or a similar spreadsheet program running on a Linux, PC, or Mac® platform.

2.3. Data Sets

Processed ESTs in a FASTA format ***(3)*** (see also [http://www.oardc.ohio-state.edu/phytophthora/gc.htm] for a sample input file). It is essential to remove vector sequences and to trim low quality sequences prior to processing.

3. Methods

3.1. Running GC to Count the Frequency of GCs

1. Download or transfer the program *GC* and the input file containing the ESTs to the appropriate directory in your Linux workstation (*see* **Note 1**).
2. Start the program by typing: gc.
3. At this point, you will be prompted to type the input file name and then the output file name.
4. The output file is a comma-formatted file that can be exported into Excel or a similar spreadsheet program.

3.2. Importing GC Output into Microsoft Excel

1. Open or import the output file with Microsoft Excel. The Text Import Wizard window will pop-up.
2. Select original data type: delimited.
3. Click Next.
4. Select delimiters: comma and deselect tab.
5. Click Next.
6. In data preview, assign column A to text format and the other columns to general format.
7. Click Finish.
8. The GC frequency data is now imported into the spreadsheet.

3.3. Description of Output

There are eight columns in the output file:

1. Column A: sequence ID.
2. Column B: GC content for frame 1 (based on the first base of the EST).
3. Column C: GC content for frame 2.
4. Column D: GC content for frame 3.
5. Column E: GC content for entire sequence.

6. Column F: Ratio of GC content frame 1/GC content entire sequence.
7. Column G: Ratio of GC content frame 2/GC content entire sequence.
8. Column H: Ratio of GC content frame 3/GC content entire sequence.

3.4. Identifying High GC Sequences

The table can be sorted in descending order based on column E to help identify high GC sequences:

1. Select columns A–H.
2. Select Data:Sort and sort based on column E and descending order.
3. Identify high GC sequences by scrolling down the file.
4. For oomycete–plant ESTs, we estimate that sequences with a GC content higher or equal to 53% have a 98% probability to be of pathogen origin (*see* **Note 2**).

3.5. Quality Check

A quality check can be performed by searching the high GC sequences against species-specific databases using the BLASTN algorithm *(4)* (*see* **Note 3**).

4. Notes

1. Ideally, the sequences should be generated using a robust base-calling program, such as PHRED *(5,6)*. It is essential to trim the ESTs for low quality sequences. Some EST data sets may have an overrepresentation of long stretches of As or Ts due to the polyadenylation signals in the mRNA. In such cases, these stretches need to be removed.
2. This estimate is based on the observation that for tomato, less than 2% of the ESTs have a GC content higher or equal to 53%.
3. The clear differences in GC content between plant and oomycete cDNA sequences may not occur in other pathosystems. The GC content of cDNAs from the examined organisms need to be determined in order to establish a reliable threshold for discrimination. In cases in which there are no clear difference in GC content, the hexamer counting method described by Hraber et al. *(7)* could be a valuable alternative.

Acknowledgments

Supported by the OARDC Research Enhancement Grant Program and Syngenta Biotechnology, Inc. Salaries and research support were provided by State and Federal Funds appropriated to the Ohio Agricultural Research and Development Center, The Ohio State University.

References

1. Birch, P. R. J. and Kamoun, S. (2000) Studying interaction transcriptomes: coordinated analyses of gene expression during plant-microorganism interactions, in *New Technologies for Life Sciences: A Trends Guide* (Wood, R., ed.), Elsevier Science, New York, pp. 77–82.
2. Qutob, D., Hraber, P. T., Sobral, B. W., and Gijzen, M. (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiol*. **123,** 243–254.
3. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*. **183,** 63–98.
4. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **17,** 3389–3402.
5. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. **8,** 175–185.
6. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. **8,** 186–194.
7. Hraber, P. T. and Weller, J. W. (2001) On the species of origin: diagnosing the source of symbiotic transcripts. *Genome Biol*. **2,** RESEARCH0037.

6

Computer Software to Find Genes in Plant Genomic DNA

Ramana V. Davuluri and Michael Q. Zhang

Summary

Gene finding is the most important phase of genome annotation. Eukaryotic genomes contain thousands of protein coding genes, and computational gene prediction would rapidly increase the pace of experimental confirmation of expressed genes at the bench. The purpose of this chapter is to discuss the use of different computer programs that identify protein-coding genes in large genomic sequences. We describe most commonly used gene prediction programs that are available on the World Wide Web and demonstrate the use of some of these programs by an example. We provide a list of these programs along with their Web uniform resource locators (URLs) and suggest guidelines for successful gene finding.

Key Words

gene prediction, protein-coding region, gene structure, splice sites, exons, computational gene finding

1. Introduction

The human ***(1)*** and *Arabidopsis* ***(2)*** genome projects and the advancement of sequencing technologies within the last decade are driving many other genome projects. The complete genome sequences of more than 800 organisms (many microbes, fungi, plants, and animals) are either complete or being sequenced (http://www.ncbi.nlm.nih.gov). One of the primary goals of any genome project is to provide a single continuous sequence for each of the chromosomes and demarcate the positions of all genes (**Fig. 1A**), along with the annotation of each component of a gene (**Fig. 1B**). Furthermore, recent advances in high-throughput technologies, such as genome-wide micro-array expression analysis, are starting to provide greater insights into the transcrip-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

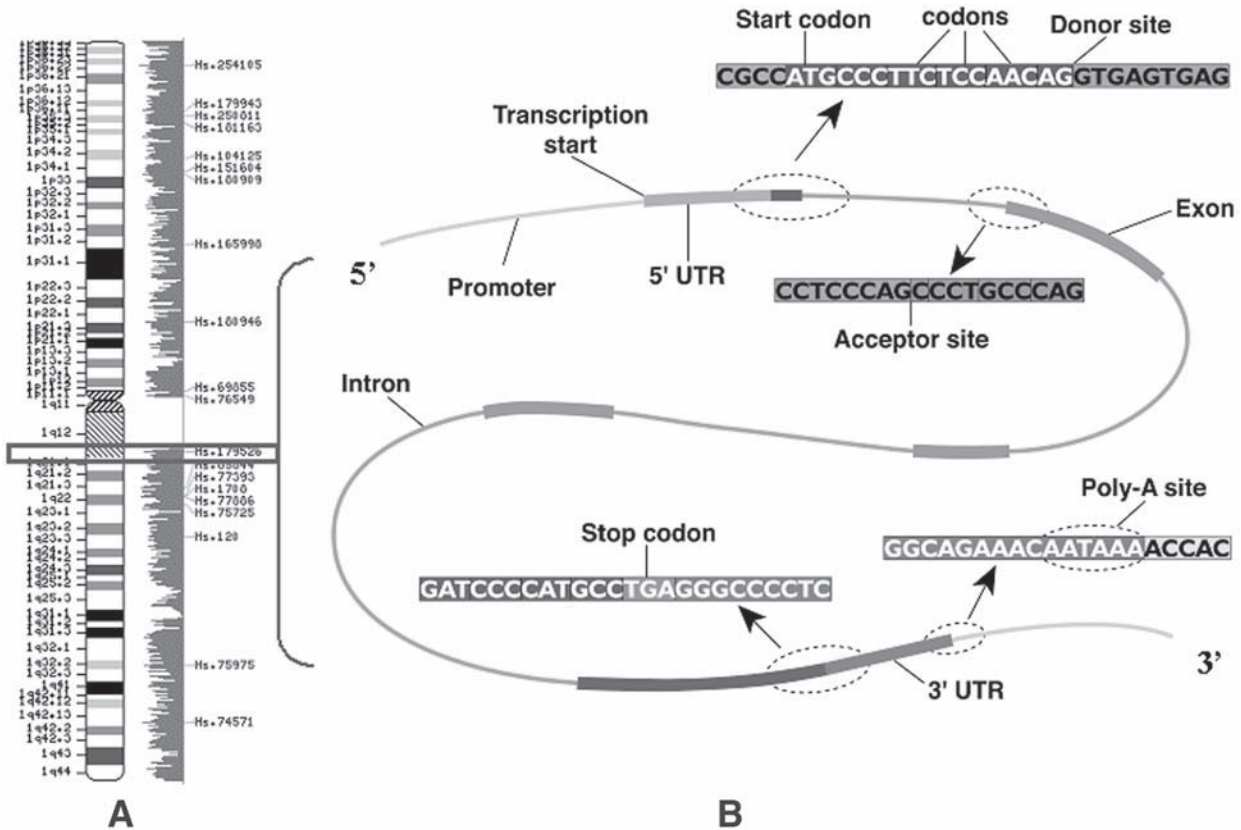


Fig. 1. Genome annotation. **(A)** Annotation of genes at chromosome level. **(B)** Annotation of individual components of a gene (such as exons, start codon, transcription start site, etc.).

tional regulation of eukaryotic cells ***(3–5)***. Integrating the genome sequence information (e.g., gene promoters) and microarray expression data would provide an initial link to functional genomics. The identification and annotation of genes at genome level will contribute to the understanding of genome-wide gene expression studies. The major focus of this chapter is to introduce different bioinformatics tools that identify genes in genomic sequences.

Gene, defined as a transcribed unit, is usually split into pieces (called exons) that are separated by intervening sequences (called introns) in the eukaryotic genomes (**Fig. 1B**). The identification of genes by computational approaches is relatively straightforward for organisms with compact genomes (such as bacteria and yeast), because exons tend to be large, and the introns are either nonexistent or short. The challenge is much greater for larger genomes (such as those of rice or maize), because the exonic "signal" is buried under nongenic "noise." In the past few years, the accuracy and reliability of computational gene finding programs have improved to a reasonable extent, such that gene predictions within a genomic region can give valuable guidance to more detailed experimental studies. Computational sequence analysis methods, which detect genes in genomic DNA, can be broadly classified into two main categories: homology-based methods, and *ab initio* methods, which we discuss in **Subheading 3**.

2. Materials

User must have access to a computer with Internet access, e.g., a personal computer (PC) running Microsoft® Windows™ or Linux, an Apple® Macintosh®, or a UNIX® workstation. The user should be familiar with the use of Netscape Navigator or Microsoft Internet Explorer. The list of commonly used gene finding and sequence alignment programs and their Web uniform resource locators (URLs) are provided in **Table 1**.

3. Methods

3.1. Gene Prediction by Homology-Based Methods

Sequence homology is a very powerful type of evidence used to detect functional elements in genomic sequences. The homology-based methods to detect genes use either intraspecies or interspecies sequence comparison in at least four different ways, as summarized below.

3.1.1. Comparison with Expressed Sequence Tags/cDNA Database

A direct comparison of a genomic sequence (query) with expressed sequence tags (ESTs) or cDNA (**Fig. 2**) can identify regions of the query sequence that correspond to processed mRNA. BLASTN ***(6)*** is a common program that iden-

Table 1
Web URLs of Gene-Prediction and Sequence Alignment Programs

Program name	Model	Organism	Web URL
AAT	MZEF+homology		http://genome.cs.mtu.edu/aat.html
BCM Search Launcher	Many gene finding programs		http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
BLAST	Sequence alignment programs		http://www.ncbi.nih.gov/BLAST
CDS (search coding region)			http://bioweb.pasteur.fr/seqanal/interfaces/cds-simple.html
Fgenesh: (**Fgenes**; **Hexon**; **TSSW**; **TSSG**; **SPL**; **Polyah**)	HMM	dicots, monocots	http://genomic.sanger.ac.uk/gf/gf.shtml http://searchlauncher.bcm.tmc.edu:9331/seq-search/gene-search.html http://www.softberry.com/nucleo.html
GeneMachine	Integrated gene finder	*Arabidopsis*	http://genome.nhgri.nih.gov/genemachine/
GeneMark.hmm	HMM	*Arabidopsis*	http://dot.imgen.bcm.tmc.edu:9331/seq-search/gene-search.html
GeneParser	DP-ANN		http://beagle.colorado.edu/~eesnyder/GeneParser.html
GeneSplicer	Marko model and MDD	*Arabidopsis*, rice	http://www.tigr.org/tdb/GeneSplicer/gene_spl.html
GeneWise2	DNA protein alignment		http://www.cbil.upenn.edu/tess/
GenLang			http://www.cbil.upenn.edu/genlang/genlang_home.html
Genomescan	HMM+protein similarity	*Arabidopsis*, maize	http://genes.mit.edu/genomescan/
Genscan	HMM	*Arabidopsis*, maize	http://genes.mit.edu/GENSCAN.html
GRAIL	ANN	*Arabidopsis*	http://compbio.ornl.gov/tools/index.shtml

Program name	Model	Organism	Web URL
MORGAN	Decision tree, HMM		http://www.tigr.org/~salzberg/
VEIL			
GLIMMER			
MZEF	QDA	*Arabidopsis*	http://www.cshl.edu/mzhanglab/
MZEF SPC	MZEF+SpliceProximalCheck		http://industry.ebi.ac.uk/~thanaraj/MZEF-SPC.html
NetGene2	ANN	*Arabidopsis*	http://www.cbs.dtu.dk/services/NetGene2/
NNSplice	ANN	Drosophila, Human, or other	http://www.fruitfly.org/seq_tools/splice.html
OrfFinder			http://www.ncbi.nlm.nih.gov/gorf/gorf.html
PredictGenes			http://cbrg.inf.ethz.ch/subsection3_1_8.html
Procrustes	Spliced alignment		http://www-hto.usc.edu/software/procrustes/index.html
PROCRUSTES	Spliced alignment program		http://www-hto.usc.edu/software/procrustes
RepeatMasker	Identifies and masks interspersed repeats		http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
RiceHMM	HMM and EST similarity	Rice	http://rgp.dna.affrc.go.jp/RiceHMM
SGP-1	Similarity based gene prediction		http://soft.ice.mpg.de/sgp-1
SIM4	Spliced alignment program		http://pbil.univ-lyon1.fr/sim4.html
SplicePredictor	Logitlinear model	*Arabidopsis*, maize	http://bioinformatics.iastate.edu/cgi-bin/sp.cgi
WebGene		*Arabidopsis*	http://www.itba.mi.cnr.it/webgene/
Xpound			ftp://igs-server.cnrs-mrs.fr/pub/Banbury/xpound/
YeastGene			http://tubic.tju.edu.cn/cgi-bin/Yeastgene.cgi

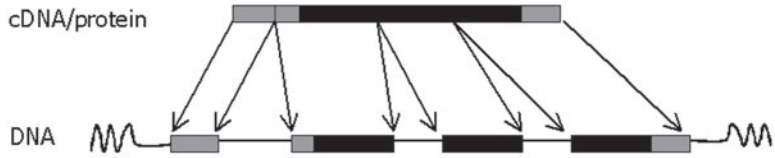


Fig. 2. Sequence alignment. Alignment of a cDNA or protein with a genomic sequence. In the cartoon showing the DNA, the rectangular boxes represent the exons, and the straight lines represent the introns.

tifies similar nucleotide sequences that exist in the databases (nr/EST) to the query sequence (*see* **Note 2**). BLASTN algorithm finds similar sequences by generating an indexed table or dictionary of short subsequences called words for both the query and the database (*see* Basic Local Alignment Search Tool [BLAST] help at [http://www.ncbi.nlm.nih.gov/BLAST] for further details). For identification of gene regions in the query sequence, choose low complexity repeat filter and select expected value as 0.1. If the query sequence is very long MegaBLAST is a better choice, as it is specifically designed to efficiently find long alignments between very similar sequences. MegaBLAST is also optimized for aligning sequences that differ slightly as a result of sequencing errors. The user can select different options. We suggest the use of expected value (e-value) of 0.1 and choose filter for low complexity repeats. When larger word size is used (default value is 28), it increases the search speed and limits the number of database hits. For BLASTN, the word size can be reduced from the default value of 11 to a minimum of 7 to increase sensitivity.

BLASTN is mainly used to pull out similar sequences from the database, and most of the times it is hard to interpret the exon boundaries. After finding a cDNA or EST match to the query sequence, one can use spliced alignment programs such as SIM4 ***(7)***, which efficiently aligns an EST or cDNA with the genomic sequence. RiceHMM ***(8)*** is another program that predicts gene domains in rice genome sequence, based on a Hidden Markov Model using a database of rice ESTs, composed of nearly 15,000 cDNAs.

3.1.2. Comparison with Protein Sequence Databases

Comparison of genomic sequence with protein sequence database by programs, such as BLASTX, can identify probable protein coding regions. Subsequently, spliced alignment programs such as Genewise ***(9)***, GeneSeqer ***(10)***, or PROCRUSTES ***(11)*** can be used to find the gene structure by comparing the genomic DNA sequence to the target protein sequences. These programs derive an optimal alignment based on sequence similarity score of the predicted gene product to the protein sequence and intrinsic splice site strength of the predicted introns.

3.1.3. Comparison of a Translated Genomic Sequence with Translated Nucleotide Database

A comparison of a translated genomic sequence with nucleotide database, which has been translated in all six reading frames, using TBLASTX can identify similarities among protein coding regions. TBLASTX can be run by selecting "Nucleotide query—Translated db [tblastx]" option from the BLAST Web page. TBLASTX takes a nucleotide query sequence, translates it in all six frames, and compares the translations to a nucleotide database (e.g., nr, est, est_human, est_others, etc.) sequences that are dynamically translated in all six frames.

3.1.4. Comparison of Genomic Sequence with Homologous Genomic Sequences from Related Species

Protein coding DNA from closely related plant species, such as sorghum and maize, show considerable sequence similarity ***(12)***. With the availability of genomes of many different organisms, comparative genomic approaches are gaining importance. VISTA/AVID ***(13)*** and PipMaker ***(14)*** can be used to compare large genomic sequences to find orthologous genomic sequences from closely related species. For example, sequence analysis of orthologous genes from rice, maize, and sorghum showed that the exons are more conserved than introns ***(12)***. The degree of sequence conservation, in terms of sequence identity, across species has been shown to be consistent with the divergence times of the respective species. The rice genes are considerably more diverged than their counterparts in maize and sorghum. For gene prediction programs, it would be best to compare two genomes that are very closely related, but distant enough that their intergenic repeat elements differ significantly. As a rule of thumb, consider two species as closely related, if those two are diverged within the last 25 million yr. For example, maize and sorghum are closely related species as they were diverged 15–20 million yr ago. If homologous genomic sequences from two species are known, then a recently developed gene prediction tool called SGP-1 ***(15)*** can be used to find protein-coding genes.

3.2. Gene Prediction by Ab Initio Methods

Homology-based methods provide useful information about gene locations as well as clues about gene function. Similarity-based methods, such as BLAST, combined with more sophisticated spliced alignment methods, such as SIM4, can give most reliable gene structure, provided there exists a full-length cDNA sequence in the database. However, most of the cDNA or EST sequences are partial, and these databases are increasing rather slowly. To help overcome these limitations, several *ab initio* gene finding programs have been

developed over the years (**Table 1**). These programs recognize signals or compositional features in an input genomic sequence by pattern matching or statistical methods. The performance of a gene finding program is typically measured in terms of the sensitivity, defined as the proportion of true signals (e.g., donor signals, exons) that are correctly predicted, and specificity, defined as the proportion of predicted signals that are correct. A program is considered accurate if its sensitivity and specificity are simultaneously high. We describe some of the most commonly used gene prediction programs trained for plant genomes. A comprehensive review of these programs can be found at Weintian Li's Bibliography on Computational Gene Recognition Web site (http://linkage.rockefeller.edu/wli/gene/). A recent review by Lincoln Stein ***(16)*** surveys the various ways the genome annotation is being carried out.

3.2.1. Splice Site Prediction Programs

Since most vertebrate, invertebrate, and plant genes have several exons; precise gene structure prediction in these organisms very much depends on the ability of splice site prediction. Many first generation gene prediction programs used simple position weight matrix methods to model the compositional biases present in the 5' and 3' splice sites. Most recent programs have investigated the correlations between different positions by using Markov models, maximal dependence decomposition models, decision tree models, and artificial neural networks. GeneSplicer, Netplantgene, Netgene2, and SplicePredictor are some of the splice site prediction programs that use splice site models. The specificity of these programs is just around 35% at a 50% sensitivity threshold in large genomic sequences ***(17)***. This is because the selection of splice sites not only depends on the strength of the splice sites but also on other factors, such as exonic and intronic enhancer signals located some distance from splice junctions ***(18)***. To get an initial assessment of potential splice sites we recommend the use of GeneSplicer ***(19)***, SplicePredictor ***(20)***, or NetGene2 ***(21)***.

3.2.2. Exon Prediction Programs

Most of the gene prediction programs have been trained to predict protein coding exons; exons corresponding to the region from translation initiation codon (ATG) to stop codon (TAA/TAG/TGA). The protein coding exons typically are of four types: *(i)* initial exons (ATG to first donor site); *(ii)* internal exons (acceptor site to donor site); *(iii)* terminal exons (acceptor site to stop codon); and *(iv)* single exons (ATG to stop codon without introns). The accuracy of splice site prediction, and hence exon prediction, by second generation programs (e.g., Genscan ***[22]***, GeneMark.hmm ***[23]***, MZEF ***[24]***, or SPL ***[25]***) is significantly higher than simple splice site prediction programs, because these programs integrate splice site models with additional types of information, such

as compositional features of exons and introns. MZEF, based on quadratic discriminant analysis, was specifically trained to predict internal exons. It was shown ***(25)*** to perform better than FGENESP, GRAIL, Genscan, and GeneMark.hmm in predicting internal exons for *Arabidopsis* genome. For predicting initial and terminal exons, Genscan and GeneMark.hmm are the best options, even though the accuracy of predicting these exons is significantly lower than that of internal exon prediction.

3.2.3. Gene Modeling Programs

The accuracy of individual exon prediction further increases by combining the reading frame compatibility of adjacent exons to make a full coding transcript. Probabilistic models, such as Hidden Markov Models, have been used to incorporate this information in Genscan and GeneMark.hmm, which model different states (exon, intron, intergenic region, etc.) of a gene. In gene modeling and predicting multiple genes in large genomic contigs, Genscan and GeneMark.hmm were shown to give comparable results and by far the best available programs for plant genomes ***(25)***.

3.3. Gene Prediction by Integrated Methods

Gene prediction by homology-based methods is perhaps the most efficient way of finding genes in genomic sequences, since the evidence of support (mRNA, EST, protein) was already derived experimentally. On the other hand, *ab initio* gene-prediction programs miss some known genes (false negatives) and predict some that are not real (false positives). Traditionally, *ab initio* gene prediction programs and homology-based approaches were used independently and combined later manually by an annotator. This process has been automated in recent programs, such as Genomescan ***(27)*** and RiceGAAS ***(8)*** that combine gene predictions with similarity comparisons to produce more reliable predictions of protein-coding regions. GenomeScan incorporates protein homology information (BLASTX hits) with the exon–intron predictions of Genscan. The input to this program consists of a genomic sequence, a selection of appropriate organism (from vertebrate, *Arabidopsis*, and maize), and a set of protein sequences (in fasta format), which may be similar to the genomic sequence. GenomeScan first masks the interspersed repetitive elements in the genomic sequence with RepeatMasker and then combines the Genscan predicted peptides with BLASTX hits. The program determines the most likely "parse" (gene structure), conditional on the given similarity information under a probabilistic model of the gene structural and compositional properties of genomic DNA for the given organism.

RiceGAAS runs Genscan (with *Arabidopsis*, maize models), RiceHMM, MZEF (with *Arabidopsis*, model), and SplicePredictor (with *Arabidopsis*,

maize models) programs and combines these predictions with BLASTN (against MAFFRICE database) and BLASTX (against nr database) homology comparisons. It also masks the repeats of *Arabidopsis thaliana* repeats by using RepeatMasker program. For RiceGAAS, the input is the genome sequence to be analyzed, which can be pasted in a window or uploaded from a file (as fasta format).

3.4. Worked Example

We discussed various gene-finding strategies in the previous sections. Now let us discuss which programs to choose and how to use those programs in a real practical scenario. Given a large genomic sequence, we suggest the following steps in arriving at probable exons that the sequence may contain.

1. Blast the sequence against nr and EST databases by using BLASTN (Megablast in case of very long sequence) program. Note the list of accession numbers of cDNAs or ESTs with "% identity" score ≥99, from the blast output.
2. Use SIM4 program to align each of the cDNA/ESTs with the genomic sequence so as to identify exons with canonical splice sites.
3. Blast the sequence against nr database by using the BLASTX program. From the output, note down the BLASTX matches that may belong to genes.
4. Submit the sequence to at least 4 different gene prediction programs and select the consensus predictions (exons). We consider a prediction as consensus prediction if it is predicted by at least half of the programs either fully (both ends of the predicted exons are same) or partially (there exists an overlapping region among the predicted exons).

To demonstrate the above steps, we use the genomic sequence in rice bacterial artificial chromosome (BAC) in GenBank® with Accession no. AP005190, which has not yet been annotated at the time writing of this chapter. Since the length of the sequence is very large (138,893 bp), we used Megablast to identify the homologous sequences from the GenBank. The program was run twice by choosing nr and EST databases. **Table 2** gives the list of high scoring segment pairs (HSPs) from the Megablast output. As BLAST is mainly a sequence similarity program, it helps us to identify the regions in the input sequence (query sequence) that are similar to known sequences (subject sequences) in the database. As the output suggests, it is hard to interpret the gene structure (exon–intron boundaries) from the output. Hence, we ran SIM4 program to align each of the EST/cDNA sequences (from the output of Megablast) with the genomic sequence AP005190. **Table 3** gives the list of exons inferred by combining various EST/cDNA alignments with AP005190 using SIM4.

Table 2
List of HSPs of AP005190 (Query) Against EST Database from Megablast Output

Subject ID	% Identity	Alignment length	Mismatches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
AU173904	100	375	0	0	47222	47596	87	461	0	743.9
AU173904	100	87	0	0	46592	46678	1	87	6.70E-38	173
AU173465	100	363	0	0	24137	24499	433	71	0	720.1
AU173465	100	72	0	0	25919	25990	72	1	6.00E-29	143.2
AU031146	100	313	0	0	14463	14775	138	450	9.00E-173	621
AU093845	99.4	317	1	1	14463	14778	381	697	2.00E-170	613
AU093845	100	116	0	0	13601	13716	266	381	3.30E-55	230.4
AU093845	100	75	0	0	12946	13020	194	268	9.70E-31	149.2
AU093845	100	74	0	0	12788	12861	125	198	3.80E-30	147.2
C97606	99.7	313	0	1	14463	14775	527	838	9.00E-170	611.1
C97606	100	116	0	0	13601	13716	412	527	3.30E-55	230.4
C97606	100	75	0	0	12946	13020	340	414	9.70E-31	149.2
C97606	100	74	0	0	12788	12861	271	344	3.80E-30	147.2
C73253	99.3	286	1	1	42747	43031	425	140	7.00E-152	551.6
C73253	100	142	0	0	43214	43355	142	1	1.00E-70	282
BI798584	100	267	0	0	14463	14729	252	518	3.00E-145	529.8
BI798584	99.1	116	1	0	13601	13716	137	252	8.00E-53	222.5
BF430535	100	259	0	0	105549	105807	35	293	2.00E-140	513.9
BF430535	100	112	0	0	106534	106645	473	584	8.00E-53	222.5
BF430535	100	99	0	0	106759	106857	585	683	4.60E-45	196.7
BF430535	100	65	0	0	106228	106292	354	418	9.00E-25	129.3
BF430535	100	64	0	0	106041	106104	291	354	3.50E-24	127.4
BF430535	100	60	0	0	106373	106432	414	473	8.60E-22	119.4

Table 2
Continued

Subject ID	% Identity	Alignment length	Mismatches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
D40524	99.6	235	1	0	82675	82909	235	1	8.00E-124	458.4
D40946	99.6	230	1	0	82680	82909	230	1	2.00E-121	450.5
AU090572	99.1	231	2	0	53526	53756	78	308	5.00E-119	442.6
AU163696	100	163	0	0	120870	121032	1	163	3.00E-83	323.6
AU163696	100	125	0	0	121227	121351	161	285	1.40E-60	248.3
AU183284	100	133	0	0	12464	12596	315	447	2.40E-65	264.1
AU183284	100	120	0	0	11103	11222	195	314	1.40E-57	238.4
AU183284	100	54	0	0	9806	9859	142	195	3.30E-18	107.5
AU093296	99.2	120	0	1	11103	11222	236	354	1.30E-54	228.5
AU093296	100	70	0	0	9591	9660	117	186	9.30E-28	139.3
AU093296	100	54	0	0	9806	9859	183	236	3.30E-18	107.5
AU173536	100	112	0	0	82229	82340	112	1	8.00E-53	222.5
BQ281772	100	108	0	0	120925	121032	72	179	2.00E-50	214.6
BE599115	100	108	0	0	120925	121032	85	192	2.00E-50	214.6
BE593685	100	108	0	0	120925	121032	76	183	2.00E-50	214.6
AW680979	100	108	0	0	120925	121032	63	170	2.00E-50	214.6
BG560418	99.1	108	1	0	120925	121032	85	192	4.80E-48	206.7
AU166259	100	84	0	0	29212	29295	356	439	4.10E-36	167
AU166259	100	38	0	0	28319	28356	322	359	1.20E-08	75.82
BI813425	100	79	0	0	83259	83337	466	388	4.00E-33	157.1
BM347731	100	77	0	0	120956	121032	736	660	6.20E-32	153.1
BM079469	100	77	0	0	120956	121032	615	539	6.20E-32	153.1
BI813794	100	77	0	0	83261	83337	476	400	6.20E-32	153.1
D39271	100	77	0	0	27502	27578	185	109	6.20E-32	153.1

Table 2
Continued

Subject ID	% Identity	Alignment length	Mis matches	Gap openings	Query start	Query end	Subject start	Subject end	E-value	Bit score
BI245296	100	69	0	0	120964	121032	481	413	3.70E-27	137.3
BI813113	100	64	0	0	83274	83337	549	486	3.50E-24	127.4
BE643512	100	64	0	0	120969	121032	1	64	3.50E-24	127.4
AU082326	100	63	0	0	133964	134026	69	131	1.40E-23	125.4
BE593268	100	60	0	0	120973	121032	1	60	8.60E-22	119.4
BJ450012	100	59	0	0	48886	48944	9	67	3.40E-21	117.5
BQ667839	100	49	0	0	120984	121032	393	345	3.20E-15	97.63
BF292448	100	44	0	0	120986	121029	1	44	3.00E-12	87.72
BF145477	100	44	0	0	120986	121029	1	44	3.00E-12	87.72
BM368889	100	43	0	0	120987	121029	1	43	1.20E-11	85.73
BE639720	100	43	0	0	120990	121032	1	43	1.20E-11	85.73
BE426858	100	43	0	0	120987	121029	1	43	1.20E-11	85.73
BQ608952	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606868	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606799	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BQ606785	100	41	0	0	120989	121029	23	63	1.90E-10	81.77
BJ321890	100	41	0	0	120989	121029	809	769	1.90E-10	81.77
BJ210114	100	41	0	0	120989	121029	62	102	1.90E-10	81.77
BI125789	100	41	0	0	120992	121032	187	227	1.90E-10	81.77
BG313503	100	41	0	0	120989	121029	24	64	1.90E-10	81.77

Table 3
List of Exons Derived from the Alignments of EST/cDNAs with AP005190 by Using SIM4

Gene no.	Exon no.	Strand	Exon begin—exon end	Supported EST/cDNA
1	1	+	*9475–9658	AU093296, AU183284
	2	+	9808–9859	AU093296, AU183284
	3	+	11104–11222	AU093296, AU183284
	4	+	12464–12704	AU093296, AU183284, AU093845, C97606
	5	+	12790–12857	AU093845, C97606, BI798584
	6	+	12947–13019	AU093845, C97606, AU031146, BI798584, AY072931
	7	+	13603–13715	AU093845, C97606, AU031146, BI798584, AY072931
	8	+	14463–14778*	AU093845, C97606, AU031146, BI798584, AY072931
2	2	–	24499–24137	AU173465
	1	–	25990–25921	AU173465
3	3	–	27370–27214	D39271
	2	–	27577–27502	D39271
	1	–	27787–27678	D39271
3	1	+	27998–28354	AU166259
	2	+	29214–29295	AU166259
5	2	–	43029–42747	C73253
	1	–	43355–43215	C73253
6	1	+	46592–46677	AU173904
	2	+	47222–47596*	AU173904
	3	+	*48878–48950	BJ450012
	4	+	49354–49793*	BJ450012
7	1	+	*53449–53756*	AU090572
8	1	+	*81944–82003	AU173536
	2	+	82219–82340*	AU173536
	3	+	*82407–82909	D40524, D40946
	4	+	83253–83711	BI813425, BI813794
9	7	–	90106–90089*	BF430535
	6	–	105804–105550	BF430535
	5	–	106103–106041	BF430535
	4	–	106290–106228	BF430535
	3	–	106432–106376	BF430535
	2	–	106645–106535	BF430535
	1	–	*106857–106759	BF430535
10	1	+	*120870–121031	AU163696, BQ281772, BG560418
	2	+	121229–121436	AU163696, BQ281772, BG560418

Table 3
Continued

Gene no.	Exon no.	Strand	Exon begin—exon end	Supported EST/cDNA
	3	+	121560–121626	BQ281772, BG560418
	4	+	122609–122625*	BQ281772, BG560418
11	1	+	*133895–134146	AU082326
	2	+	134200–134215*	AU082326

*Might be an incomplete exon due to partial EST/cDNA.

Table 4
List of HSPs of AP005190 (Query) Against nr Database from BLASTX Output

Subject ID	% Identity	Alignment length	Subject start	Subject end	Query start	Query end	E-value	Bit score
AAC19401	27%	212	225	376	16622	15987	4e-24	189
AAC19401	42%	69	371	439	15925	15719	4e-24	62.8
AAC19401	41%	51	66	116	18173	18021	0.11	44.7
AAC19401	38%	39	155	193	17145	17029	0.11	42.4
AAB17501	30%	213	223	377	16625	15987	2e-25	88.6
AAB17501	38%	70	372	441	15925	15716	2e-25	55.8
AAB17501	42%	50	66	115	18170	18021	1e-06	47.0
AAB17501	37%	37	122	158	17318	17208	7e-05	40.0
AAB17501	30%	36	157	192	17136	17029	7e-05	34.7
AAB17501	41%	31	32	62	18360	18268	1e-06	33.5
AAD27547	97%	1520	1	1520	62266	66825	0	2915
AAM08795	98%	1520	265	1784	62266	66825	0	2942
AAM08795	98%	203	1	203	61125	61733	1e-113	414
AAK92543	97%	1520	194	1713	62266	66825	0	2929
AAK92543	97%	140	1	140	61314	61733	7e-73	281
BAB86564	98%	1100	1	1100	86635	83336	0	2175
AAD19359	32%	1065	832	1876	119222	116118	1e-129	466

Next, we ran "Nucleotide query—Protein db [BLASTX]" program. Select "TRANSLATED query—PROTEIN database [BLASTX]" for Choose a translation options and nr for database options. Since the sequence is very long, we submitted the sequence as three pieces (1–50 K, 50–100 K, and 100 K to rest) to save running time, which was done by entering corresponding values of each subsequence in "from" and "to" windows of Set subsequence options. The rest of the values were left as default. **Table 4** gives the list of HSPs from

Table 5
List of Consensus Exons Predicted by at Least Two Gene-Prediction Programs in the Genomic Sequence with Accession No. AP005190

Strand	Type	Ex. Begin–Ex. End	Programs predicted
+	Intr	370–459	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Intr	668–712	Genscan (A), GeneMark.hmm (M)
+	Intr	802–872	Genscan (A), GeneMark.hmm (M)
+	Intr	1501–1633	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	1945–2033	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Term	4279–4049	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Init	5382–5320	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Init	8153–8162	Genscan (A), Genscan (M)
+	Intr	9743–9859	Genscan (A), Genscan (M)
+	Intr	12464–12704	Genscan (A), GeneMark.hmm (M)
+	Intr	12790–12857	Genscan (A), GeneMark.hmm (M)
+	Intr	12947–13019	GeneMark.hmm (M), Mzef (A)
+	Intr	13603–13715	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Term	14463–14615	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	15500–15279	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	15912–15632	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	16226–16112	Genscan (A), GeneMark.hmm (M)
–	Intr	16634–16347	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	16829–16779	Genscan (A), GeneMark.hmm (M)
–	Intr	18173–18003	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	20200–19268	Genscan (A), Genscan (M)
–	Term	24499–24380	Genscan (A), GeneMark.hmm (M)
–	Intr	25684–25613	Genscan (A), GeneMark.hmm (M)
–	Intr	25997–25921	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	27571–27141	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	29214–29427	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	30478–30644	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	31529–31653	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	32807–32902	Genscan (A), GeneMark.hmm (M)
+	Intr	32961–33009	GeneMark.hmm (M), Mzef (A)
+	Intr	33144–33198	Genscan (A), GeneMark.hmm (M)
+	Intr	39059–39180	Genscan (A), Genscan (M)
+	Term	41035–41106	Genscan (A), Genscan (M)
+	Init	43393–43699	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	44245–44360	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	44447–44535	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)

Table 5
Continued

Strand	Type	Ex. Begin–Ex. End	Programs predicted
+	Intr	45293–45338	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	46050–46218	Genscan (A), Mzef (A)
+	Intr	46595–46677	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	47222–47602	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	48259–48950	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	49354–49909	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	50151–50468	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Term	50751–50795	Genscan (M), GeneMark.hmm (M)
–	Term	53795–53682	Genscan (A), GeneMark.hmm (M)
–	Intr	53973–53875	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	54140–54068	Genscan (A), GeneMark.hmm (M)
–	Intr	54335–54225	Genscan (A), GeneMark.hmm (M)
–	Intr	54605–54432	Genscan (A), GeneMark.hmm (M)
–	Intr	55400–54715	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	55547–55402	Genscan (A), Genscan (M)
–	Intr	55814–55673	Genscan (A), Genscan (M)
–	Intr	57329–55889	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
–	Init	58233–57914	Genscan (A), Genscan (M)
+	Init	60906–60917	Genscan (A), Genscan (M)
+	Intr	61125–61718	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	62266–66693	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	67890–67955	Genscan (A), GeneMark.hmm (M)
+	Intr	68046–68188	Genscan (A), GeneMark.hmm (M)
+	Intr	69099–69391	Genscan (A), GeneMark.hmm (M)
+	Intr	72191–73594	Genscan (A), GeneMark.hmm (M)
+	Term	73703–73858	Genscan (A), GeneMark.hmm (M)
–	Intr	82264–82166	Genscan (A), Genscan (M)
–	Intr	86635–83343	Genscan (A), Genscan (M)
+	Init	94228–94246	Genscan (A), Mzef (A)
–	Sngl	98915–97443	Genscan (A), Genscan (M)
+	Intr	103554–103766	Genscan (A), Genscan (M)
–	Intr	10103–106041	GeneMark.hmm (M), Mzef (A)
–	Intr	106290–106228	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Intr	106432–106376	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Intr	106645–106535	Genscan (A), Genscan (M), GeneMark.hmm (M)
–	Init	107034–106759	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Intr	112457–112600	Genscan (A), GeneMark.hmm (M)

Table 5
Continued

Strand	Type	Ex. Begin–Ex. End	Programs predicted
+	Intr	112696–113452	Genscan (A), GeneMark.hmm (M)
+	Intr	113495–114083	Genscan (A), GeneMark.hmm (M)
+	Intr	114248–114667	Genscan (A), Genscan (M), GeneMark.hmm (M), Mzef (A)
+	Intr	114743–114802	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Term	115053–115739	Genscan (A) GeneMark.hmm (M)
–	Term	118976–116094	Genscan (A), GeneMark.hmm (M)
–	Init	119460–119294	Genscan (A), GeneMark.hmm (M)
+	Init	120929–121031	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Intr	121229–121436	Genscan (A), GeneMark.hmm (M), Mzef (A)
+	Term	121560–121680	Genscan (A), GeneMark.hmm (M), Mzef (A)
–	Term	126660–126599	Genscan (A), GeneMark.hmm (M)
–	Intr	126961–126811	Genscan (A), GeneMark.hmm (M)
–	Init	127447–127307	Genscan (A), Genscan (M), GeneMark.hmm (M)
+	Init	129895–131341	Genscan (A), GeneMark.hmm (M)
+	Intr	132275–132331	Genscan (A), Mzef (A)
+	Intr	133577–133610	Genscan (A), Genscan (M)

In the column headings: type stands for type of exon; *Init, Intr,* and *Term* stand for *Initial, Internal, and terminal* exons, respectively, and ex. stands for exon.

BLASTX output. The values in columns query start and query end would give the regions in the genomic sequence AP005190 that may belong to probable genes.

Finally, we submitted the genomic sequence AP005190 to four gene-finding programs Genscan with *Arabidopsis* model, Genscan with maize model, GeneMark.hmm with rice model, and MZEF with *Arabidopsis* model. Default values were selected for other parameters for each of the programs used. As none of the programs is good enough to predict the complete gene structure, we considered only the exon predictions. We compiled the list of all consensus exons that were predicted by at least two programs. We consider an exon as a consensus prediction if there exists an overlapping region among the predictions of at least two different programs. **Table 5** gives the list of all such exons.

4. Notes

1. Despite great progress, gene prediction by computational approaches alone is still far from perfect. The existing programs have reached a reasonable sophisti-

cation in identifying >90% of the nucleotides in a given genome as coding or noncoding (Stormo, 2000). We suggest using computational tools to identify a nucleotide as either coding or noncoding. But, identifying the exact boundaries of all the exons and assembly of the exons into different genes might be much harder and is not possible by computational approaches alone. However, even the partial predictions are of immense value to design the experiments that can determine the complete gene structure faster than would be possible by experimental methods alone.

2. Similarity-based methods (e.g., BLASTN, BLASTX) are perhaps the best to determine a given region of the genome is transcribed or not. A BLASTN match to a cDNA/EST or BLASTX match to a protein is good evidence that the region belongs to a gene. However, these methods have their own limitations. Most of the cDNAs or ESTs are incomplete and may contain one or more introns, which could lead to misclassification of intron region as exon. Some cDNA sequences may contain repetitive elements that will cause false genomic matches. Protein databases may contain potentially incorrect predicted proteins. BLASTX matches to predicted protein sequences should be avoided. Partial BLASTX alignment to a target protein should not be considered, as the protein may not be a true ortholog of the source gene and only shares some domains. We should note that the similarity data (cDNA/EST data) is never complete. Even the most comprehensive cDNA projects will miss low copy number transcripts and those transcripts whose expression is low, cell- or tissue-specific, or expressed only under unusual conditions.
3. Almost all gene finding programs can predict only protein coding regions and have not been trained to predict untranslated exons and untranslated portion of first and last coding exons.
4. Before running any gene-finding program, we suggest the use of programs such as RepeatMasker, which identifies known classes of interspersed repeats, and LINEs and SINEs, which exist in noncoding regions of the genome.
5. Most of the gene finding programs are based on statistical pattern recognition methods that require a training data. This makes the program organism-specific depending on the training data. So, while running a gene prediction program, select the organism of the genomic sequence. If the program was not trained on the organism of your choice, select the most closely related one. If the genome of your choice does not exist and has low gene density, then there may be more false positive predictions by choosing another genome with high gene density.

References

1. Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
2. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.
3. Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728.

4. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297.
5. Finkelstei, D., Ewing, R., Gollub, J., Sterky, F., Cherry, J. M., and Somerville, S. (2002) Microarray data quality analysis: lessons from the AFGC project. *Arabidopsis* Functional Genomics Consortium. *Plant Mol. Biol.* **48,** 119–131.
6. Altschul, S. F.,Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
7. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8,** 967–974.
8. Sakata, K., Nagamura, Y., Numa, H., et al.. (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30,** 98–102.
9. Birney, E. and Durbin, R. (2000) Using GeneWise in the Drosophila annotation experiment. *Genome Res*. **10,** 547–548.
10. Usuka, J., Zhu, W., and Brendel, V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16,** 203–211.
11. Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*, **93,** 9061–9066.
12. Schmidt, R. (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. *Plant Mol. Biol.* **48,** 21–37.
13. Mayor, C., Brudno, M., Schwartz, J. R., et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047.
14. Schwartz, S., Zhang, Z., Frazer, K. A., et al. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10,** 577–586.
15. Wiehe, T., Gebauer-Jung, S., Mitchell-Olds, T., and Guigo, R. (2001) SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res.* **11,** 1574–1583.
16. Stein, L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2,** 493–503.
17. Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354.
18. Berget, S. M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270,** 2411–2414.
19. Pertea, M., Lin, X., and Salzberg, S. L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29,** 1185–1190.
20. Brendel, V. and Kleffe, J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.* **26,** 4748–4757.
21. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. (1996) Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3439–3452.
22. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

23. Lukashin, A. V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115.
24. Zhang, M. Q. (1998) Identification of protein-coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. *Plant Mol. Biol.* **37,** 803–806.
25. Solovyev V. V., Salamov A. A., and Lawrence C. B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22,** 5156–5163.
26. Pavy, N., Rombauts, S., Dehais, P., et al. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*. **15,** 887–899.
27. Yeh, R. F., Lim, L. P., and Burge, C. B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816.

7

Genomic Colinearity as a Tool for Plant Gene Isolation

Wusirika Ramakrishna and Jeffrey L. Bennetzen

Summary

Plant genomes show genomic colinearity in spite of the tremendous variability exhibited in their genome size and chromosomal constitution. Comparative genetics can assist in isolation of a mapped gene in a large genome plant species using a small genome plant as a surrogate. Here, we describe various steps involved in the process of gene isolation using genomic colinearity. This involves fine resolution mapping in the large genome species and using common low copy number DNA markers that map to orthologous regions in small and large genome species to isolate candidate genes from the small genome species. Further, alternate strategies are described in cases where the targeted gene is absent in the orthologous region of the small genome species. We also discuss various technologies that can be used for the confirmation of candidate genes.

Key Words

colinearity, comparative mapping, comparative sequence analysis, functional analysis, gene isolation, genomic sequencing, rearrangement

1. Introduction

Plant genomes vary tremendously in genome size, chromosome number, and chromosome morphology ***(1,2)***. In spite of the great diversity observed among plant genomes, significant genomic colinearity has been revealed by comparative genetic mapping ***(3,4)***. Most of the observed macro and microcolinearity among plant genomes is limited to low copy number DNA probes, primarily genes. The large genome sizes of important crop plants, such as barley, maize, and wheat, can make map-based cloning extremely difficult in these species. Hence, it may be easier to use comparative maps to isolate a mapped gene from a large genome using a related plant with a small genome.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

Markers linked to the gene of interest and prior knowledge about colinearity of this region between large and small genomes are essential to isolate a gene using this approach.

The use of low copy number DNA markers as restriction fragment-length polymorphism (RFLP) probes resulted in the generation of genetic linkage maps of many important plant species. Comparative genetic maps based on RFLP probes revealed extensive colinearity and conservation of gene order and content among closely related plant species. This is especially striking for the grass genomes (for instance barley, maize, rice, sorghum, and wheat) that diverged from a common ancestor 50–70 million yr ago (mya). Further, quantitative trait loci (QTL) controlling important agronomic traits were also mapped to colinear regions among grass genomes ***(5)***. The frequency of major chromosomal rearrangements observed among genomes depends on the degree of relatedness of the species investigated. In general, more closely related species show fewer rearrangements, but there are notable exceptions ***(6,7)***. This indicates that chromosomes in some lineages are less stable than others.

In dicotyledonous plants, colinearity has been observed in the Solanaceae (e.g., pepper, tomato, and potato), the Brassicaceae and some leguminous species ***(8–11)***. Despite the general colinearity exhibited by comparative genetic maps, rearrangements that involve regions smaller than a few centiMorgans may occur and would be missed by most recombinational mapping studies. Comparative sequence analysis involving large genomic segments can detect these rearrangements. In the grass genomes, investigations of microcolinearity have been limited to a few regions ***(12,13)***. These sequencing studies have revealed small rearrangements, including deletions, duplications, inversions, and translocations of small gene blocks ***(14–19)***. In dicots, comparative genomic structure analyses with DNA sequences are mainly limited to comparisons of Brassica species and tomato to the completely sequenced *Arabidopsis* genome ***(20–22)***. However, the ancestral *Arabidopsis* genome has undergone a high frequency of chromosomal mutation and, thus, extensive genomic rearrangement relative to distantly related dicots like soybean and tomato ***(23–26)***. This complicates comparative genomic analysis using *Arabidopsis*. However, the general conservation of gene content between *Arabidopsis* and most other plants ***(27)*** often allows the use of *Arabidopsis* as a surrogate for gene isolation in different plant species. Presumed orthologues of several *Arabidopsis* genes have been cloned from cereal genomes using this approach **(28)**.

Conservation of gene content and gene order among closely related plant species greatly assists in gene identification and annotation. Even in closely related plant genomes, whose ancestors diverged from each other <10 mya, only genes are conserved in orthologous regions. All of the plant species with large genomes studied to date have been invaded by retrotransposons within

the last 6 million yr ***(29,30)***, and these sequences vary greatly between species. Other sequences between genes also evolve rapidly ***(14,16,18,19,31)***. Hence, plant species that diverged from each other >50 mya only have exonic regions conserved among genes. This feature has been used to improve gene annotation with great success ***(16,18,19)***. Gene structure can be predicted more accurately using comparative sequence analysis than by the combined use of expressed sequence tags (ESTs), homology to entries in protein databases, and gene prediction programs ***(18,19)***. Conservation of genomic colinearity, gene content, and order among plant genomes separated by less than 100 million yr greatly assists in gene isolation from cross-species comparisons.

Differences in gene content are sometimes observed in otherwise microcolinear regions of plant genomes ***(16,17,19,32)***. This phenomenon can complicate gene isolation, but does not completely invalidate the approach. Under almost all circumstances, a small genome species will provide numerous DNA markers on a single bacterial artificial chromosome (BAC), which permits more detailed mapping in the large genome species. Chromosome walking involves identifying low copy number DNA markers that are tightly linked to the gene of interest and using them as probes to screen large insert BAC libraries to identify appropriate clones. Repeated rounds of such screening using low copy number regions from a series of BACs may be required to identify overlapping clones extending toward the targeted gene. Chromosome walking is often difficult with large genomes such as barley, maize, and wheat. In these cases, related plant species with small genomes such as rice, which show genomic colinearity with the large genome species, can be used to identify and isolate the desired gene. This approach has potential pitfalls, especially with respect to some disease resistance genes ***(33–35)***. Resistance gene regions often undergo rapid rearrangement that results in a lack of microcolinearity caused by deletion or translocation of the targeted loci. However, at the very least, the comparative genomic approach provides numerous probes from one species, which can be used for gene mapping and isolation in another species.

As physical maps become available, more accurate, and more detailed for large genome species, the need for a small genome surrogate diminishes as a map-based cloning tool. However, thousands of large genome plant species have genes for important traits that have been mapped or can be mapped in a comparative mode. Hence, these less-studied species will continue to have use for a small genome surrogate. In this chapter, we describe methods for plant gene isolation based on comparative genetic map and/or genomic sequence information. This technique involves identification of colinear regions, followed by clone selection, and finally, sequence analyses to identify the gene of interest.

2. Materials

2.1. General

1. High density genetic linkage maps of plant species that show colinearity and from which the gene of interest is to be cloned. Low copy number DNA markers tightly linked to the locus must be identified using a mapping population developed from a cross between two parents polymorphic for the gene of interest. It is best if the mapping population is large, involving recombination through at least 200 (preferably >1000) recombinant meioses. Colinear genetic linkage maps for different plant species can be found at (www.gramene.org), (www.agron.missouri.edu/maps.html), (www.arabidopsis.org), and (www.sgn.cornell.edu/maps/tomato_arabidopsis_map.html).
2. BAC library filters (often available from [http://www.genome.clemson.edu] and [http://hbz.tamu.edu/bacindex.html]).

2.2. Standard Reagents and Buffers

1. All-in-one random prime labeling mixture (Sigma).
2. 20× Sodium chloride sodium phosphate EDTA (SSPE).
3. 20% Sodium dodecyl sulfate (SDS).
4. α-[^{32}P] dCTP (3000 Ci/mmol) (Amersham Pharmacia Biotech).
5. Hybridization oven (Hybaid).
6. Restriction enzymes.
7. Tris-borate-EDTA (TBE) buffer.
8. Agarose (Invitrogen).
9. Horizontal gel electrophoresis unit (Maxicell).
10. Power inverter (MJ Research).
11. DNA size markers, midrange II pulsed-field gel (PFG) marker (New England Biolabs), high molecular weight (Invitrogen), and 1-kb DNA ladders (Invitrogen).
12. Nylon membrane, Hybond® N (Amersham Pharmacia Biotech).
13. Large construct kit (Qiagen).
14. 10 m*M* Tris-HCl, pH 8.5.
15. Hydroshear device (GeneMachines).
16. Mung bean nuclease (Amersham Pharmacia Biotech).
17. 25:24:1 Phenol:chloroform:isoamyl alcohol mixture (Sigma).
18. Isopropanol.
19. Shrimp alkaline phosphatase (Roche Molecular Biochemicals).
20. *Taq* DNA polymerase (Promega).
21. DarkReader™ (Clare Chemical Research).
22. QIAex® II gel extraction kit (Qiagen).
23. TOPO® TA cloning kit for sequencing (Invitrogen).
24. DH10B electroMAX cells (Invitrogen).
25. Glass beads (Fisher Scientific).
26. Qpix colony picker (Genetix).

27. 384-Well culture trays (Genetix).
28. REAL prep 96 plasmid kits (Qiagen).
29. DNA sequencing kits, big dye terminator v3.0 cycle sequencing ready reaction kit (Applied Biosystems); dGTP BigDye™ terminator ready reaction kit (Applied Biosystems); dRhodamine terminator cycle sequencing ready reaction kit (Applied Biosystems); DYEnamic ET terminator cycle sequencing kit (Amersham Pharmacia Biotech).
30. ABI PRISM™ 3700 DNA analyzer (Applied Biosystems).
31. Squeeky-Clean 96-well column plates (Bio-Rad).
32. Thermofidelase I (Fidelity Systems).
33. EZ::TN <TET-1> insertion kit (Epicentre Technologies).

2.3. Software and Computer

1. Information about PHRED (base caller), PHRAP (assembler) and CONSED (graphical editor) is available at (http://www.phrap.org).
2. Gene prediction programs are available at (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi), (http://genes.mit.edu/GENSCAN.html), and (http://www.softberry.com/berry.phtml).
3. The Web site for GenBank® is (http://www.ncbi.nlm.nih.gov) and that of GeneSeqer is (http://gremlin3.zool.iastate.edu/cgi-bin/prg/gs.cgi). The Institute for Genomic Research (TIGR) rice repeat database is available at (http://www.tigr.org/tdb/rice/blastsearch.shtml).

3. Methods

3.1. Identification of Colinear Regions

1. The genetic map position of the targeted locus in the plant species with a large genome size must be determined accurately by segregation analysis of the locus with tightly linked markers. These markers should map to a colinear region in the plant species with a small genome to enable isolation of the targeted locus. Comparative genetic linkage maps with common RFLP markers serve as the best starting point. The maize genome is about 2400 Mb in size, corresponding to a genetic map of about 2500 cM ***(36)***. This translates to an average of 1 Mb/cM for the maize genome. A large mapping population of 5000 gametes with no recombinants in the segregating progeny makes it likely that the targeted gene is present within a 500-kb region. However, different regions of the genome can vary significantly in their recombination frequencies. For instance, 1 cM may be dozens of Mb near paracentromeric heterochromatin or <100 kb in a gene-rich region. The rice genome has a size of 450 Mb and a genetic map of about 1600 cM (http://rgp.dna.affrc.go.jp/ine.pl). This makes map-based gene isolation much easier in rice than in maize.
2. In cases where the gene of interest is absent in the small genome ***(37,38)***, we can use markers from the orthologous region in the smaller genome to fine-map in the larger genome. The nearly complete rice genomic sequence provides abun-

dant information for choosing suitable probes. The maize BAC libraries are screened with suitable probes to identify BACs that harbor the gene of interest.
3. The next step is to look for the presence of flanking markers (tightly linked to the targeted gene) on contiguous BACs in maize. The results of such studies will show whether overall colinearity is maintained in the region (*see* **Note 1**).

3.2. Clone Selection and Mapping

1. Several thousand clones from the small genome BAC library are screened for individual clones that show homology to DNA markers mapped in the colinear regions in different plant species.
2. Labeled probes are prepared using all-in-one random prime labeling mixture as per manufacturer's instructions.
3. BAC library filters are hybridized with labeled probes using 5× SSPE and 7.5% SDS at a suitable temperature (55°–65°C).
4. The filters are washed twice for 15 min each with 2× SSPE and 0.1% SDS. This is followed by a wash at the hybridization temperature with 1× SSPE and 0.1% SDS.
5. The filters are then exposed to X-ray films.
6. Positive BACs found in the BAC libraries are individually digested with restriction enzymes with 8-bp specificities such as *Asc*I, *Not*I, *Pac*I, *Pme*I, and *Swa*I.
7. Restriction fragments are separated by field inversion gel electrophoresis in 0.8% agarose gels in 1× TBE buffer. The gels are run at 4°C for 14 h at 200 V using program 2 on a programmable power inverter (*see* **Note 2**). Depending on the size of the restriction fragments to be resolved on the gel, different size standards can be used that include midrange II PFG marker, high molecular weight, and 1-kb DNA ladders.
8. Fragments observed on the agarose gel are compared between BACs to identify common fragments. All possible single and double digestions are analyzed with the restriction enzymes with one or more sites within the BACs. The BACs will form a single contiguous array (contig) if the probe hybridizes to only one region in the genome. BAC libraries of polyploid genomes generally show BACs organized in more than one contig.
9. The DNA is transferred to nylon membranes and hybridized with suitable probes to confirm that all the BACs are from the same locus. This also helps confirm and orient the BAC contig.

3.3. Construction of Shotgun Libraries

1. DNA from BACs is extracted using the large construct kit. The DNA is dissolved in 120 μL of sterile water or 10 mM Tris-HCl, pH 8.5, and sheared with a Hydroshear device to a size range of 4–8 kb as per manufacturer's instructions (*see* **Note 3**).
2. The sheared fragments are converted to blunt-ended fragments with mung bean nuclease in a total vol of 50 μL at 37°C for 20 min. The DNA is extracted with a phenol–chloroform–isoamyl alcohol mixture and precipitated with isopropanol.

3. The DNA is dissolved in 40 μL of sterile water and dephosphorylated with shrimp alkaline phosphatase in a total vol of 50 μL at 37°C for 1 h.
4. "A" tails are added by incubation with *Taq* DNA polymerase and dATP at 72°C for 30 min.
5. The DNA is run through a 1% agarose gel under DNase-free conditions (*see* **Note 4**). It is important not to expose the gel to UV-light, since it dramatically reduces the number of clones obtained from the library. We use DarkReader to view the gel and excise the agarose gel piece with the desired range of DNA fragments.
6. The DNA is eluted in a small vol (6 μL) from the gel using the QIAex II gel extraction kit.
7. These fragments are cloned in the vector pCR4-TOPO using the TOPO TA cloning kit for sequencing, following the instructions of the manufacturer.
8. The resulting DNA is transformed into DH10B electroMAX cells by electroporation.
9. The cells are plated on 25 μg/mL kanamycin plates (Genetix) using glass beads to improve colony dispersal.
10. The plates are incubated at 37°C for 16 h.
11. Colonies are picked using a Qpix colony picker into 384-well culture trays filled with 60 μL of terrific broth culture medium plus 8% glycerol. After overnight growth (14–18 h) at 37°C, cultures are frozen at –80°C until needed.

3.4. Sequencing

1. REAL prep 96 plasmid kits are used to prepare DNA minipreps from 1.3 mL cultures grown in deep 96-well plates for 14–18 h at 37°C with shaking at 300 rpm. DNA is resuspended in 50 μL of water, with 4 μL used for each sequencing reaction.
2. Clones are sequenced from both directions using big dye terminator chemistry and run on an ABI 3700 capillary sequencer after terminator clean-up using Squeeky-Clean 96-well column plates.
3. Base calling and quality assessment are done using PHRED ***(39)***. Contiguous sequences (contigs) are assembled by PHRAP once the coverage has reached 8–12×, and the sequences are edited with CONSED ***(40)***. The final error rate is estimated using CONSED.
4. Sequence coverage of 3–5× can generally identify candidate genes in the BAC for further analysis. However, at this stage there are many contigs, and some genes may not be present as full length in one contig. Also, if there are two RFLP markers flanking a gene of interest, a rough draft may not give the exact location of candidate genes relative to markers in the BAC or BAC contig (i.e., whether candidate genes are present between the flanking markers or lie outside the markers). Completing BAC sequences is, thus, very useful (although expensive), because it shows the precise location of all candidate genes in the BAC or BAC contig.
5. To sequence the BAC completely, gaps are closed by a combination of different

approaches, including the use of different sequence chemistries, the thermofidelase enzyme, polymerase chain reaction (PCR) amplification of gaps, shotgun sequencing of transposon-inserted subclones that span a gap, and direct sequencing of BAC template ***(19)***. For different chemistries, reactions are primed with custom oligonucleotides using drhodamine, Big Dye dGTP, and ET chemistries.

6. Additional large-insert (8–12 kb) shotgun libraries are constructed (*see* **Note 5**) when subclones that span gaps are not available.
7. When gaps are due to repetitive regions, subclones that either start or end in unique regions with the remaining portion in the repetitive region are assembled separately and inserted into the main assembly (*see* **Note 6**).

3.5. Sequence Analyses and Annotation

1. The first step in the sequence analysis of colinear BACs (for instance, when a colinear sorghum BAC is sequenced to isolate a gene based on the genetic map location in maize) is the delimitation of regions that are conserved and not conserved relative to rice. Conserved regions are usually or always genes, while the unconserved regions are usually not genes ***(16,18,19,31)***.
2. Complete sequences from orthologous BACs are compared using the program DOTTER ***(41)*** to identify the conserved regions (*see* **Note 7**).
3. Genes are predicted using multiple gene-finding programs such as GeneMark.hmm, GENSCAN and FGENESH (*see* **Note 8**).
4. The basic local alignment search tool (BLAST) ***(42)*** is used to perform searches of sequences from the BACs with National Center for Biotechnology Information (NCBI) expressed sequence tags database (dbEST) and nonredundant databases (www.ncbi.nlm.nih.gov/BLAST). BLASTN, BLASTX, TBLASTN, and TBLASTX algorithms are used for this purpose ***(43)***.
5. Gene structure is best determined by a combination of the gene prediction programs mentioned above, along with GeneSeqer, which generates splicing alignments of significant ESTs with the BAC genomic sequence.
6. The conserved regions are generally limited to the exons ***(14,18,31)*** (*see* **Note 9**) that encode proteins. The complete predicted cDNA and predicted protein are determined from the annotated structure of the gene. The predicted cDNA is aligned with ESTs to confirm the annotated exon–intron boundaries. The structure of the most homologous annotated gene (often from *Arabidopsis*) is then used to further refine the gene structure.
7. The criteria used to define a gene are *(i)* a match to a sequence in a protein database using BLASTX ***(43)***; *(ii)* a match to ESTs or cDNAs, or *(iii)* a prediction as a gene by two or more gene prediction programs (*see* **Note 10**). These criteria are used after excluding identified transposons.
8. The presence of transposable elements is determined by BLAST searches to the GenBank/European Molecular Biology Laboratory (EMBL) nr database and TIGR rice repeat database. In addition, homology searches to known transposable elements and sequence comparison to itself (same sequence comparison) are

done using COMPARE, REPEAT, GAP (Wisconsin Package Version 10.1; Genetics Computer Group, Madison, WI, USA), DOTTER, and cross-match (http://www.phrap.org/phrap.docs/general.html). Transposons that are not highly degenerate ***(44)*** will usually have numerous specific features that allow their identification. These include terminal inverted or direct repeats and short target site duplications.

3.6. Confirmation of Candidate Genes

The possible functions of candidate genes can be investigated using several independent approaches. Sequence analyses and annotation, as described above, using comparative sequence analyses, gene finding programs, and BLAST searches, identify putative genes. Sequence variations and gene structure analysis of the genes identified in the region, for instance in susceptible and resistant lines in case of disease resistance genes, can help verify a candidate gene. For instance, preliminary mapping, cloning, sequencing, gene finding, and BLAST searches identified two candidate genes for barley *Rpg1*. These were tested by segregation analysis in 8518 gametes and by sequence analysis in barley lines susceptible and resistant to stem-rust ***(38)***.

Additional experimental analyses can be performed to evaluate candidate gene function. Several approaches can be used, as feasible, in the plant species being investigated. These include mutation analysis and expression analysis.

3.6.1. Mutation Analysis

1. Analysis of knock-out mutations (for instance T-DNA or transposon insertions) (*see* **Note 11**) (*see* also Chapters 10 and 11).
2. Wild-type lines that either have a nonfunctional or an overexpressed gene of interest can be generated by transforming wild-type plants with antisense or sense gene constructs.
3. RNA interference (RNAi) can be employed, where homologous double-stranded RNA (dsRNA) is used to suppress a gene, generally resulting in a null phenotype (reviewed in ref. ***45***).
4. Complementation studies, where a wild-type copy of the gene of interest is transformed into the mutant to see if the T1 progeny yields wild-type phenotype and whether this trait co-segregates with the transgene in subsequent generations.
5. Searching for point mutants by targeted induced local lesions in genomes (TILLING) to provide an allelic series of mutations ***(46)*** (*see* also Chapter 15).

3.6.2. Expression Analysis

1. Tissue-specific expression of the genes can be studied using Northern analysis, microarrays, reporter constructs, or reverse transcription-polymerase chain reaction (RT-PCR) to see if the expression patterns agree with the predicted biology of the targeted gene.

4. Notes

1. This assumption is true in most cases, although it's accuracy varies between species and across different genomic regions. For instance, comparative sequence analyses of maize and rice or sorghum revealed that, on average, 80–90% of the genes are colinear. However, this strategy will not work when colinearity is disturbed by deletion or translocation of some genomic segments.
2. The program has to be adjusted according to the band resolution desired on the gel. To achieve optimal resolution, electrophoretic conditions must be standardized.
3. The speed code at which the shearing apparatus gives 4- to 8-kb fragments must be optimized prior to shearing the BAC DNA.
4. The TBE buffer should be autoclaved. The agarose gel piece should be excised using a sterile razor blade.
5. Large-insert library subclones are inoculated by hand into 96-well plates containing terrific broth and kanamycin. The colony size is small, so the colonies do not grow well when inoculated using the Qpix colony picker.
6. When repeats are almost identical, they tend to assemble in an incorrect manner. In such cases, sequences have to be assembled under high stringency. In extreme cases, manual assembly must be performed using the restriction map and large insert subclones.
7. Another useful program for sequence comparisons is Artemis Comparison Tool (ACT) available at (http://www.sanger.ac.uk/Software/ACT).
8. However, if the gene is not present in the small genome surrogate, as observed in the case of the barley stem rust-resistance gene *Rpg1* lacking an orthologue in rice ***(37)***, gene annotation using comparative sequence analysis cannot identify the gene. In this case, the identified flanking markers from rice were essential in delimiting the *rpg1* locus genetically and physically in barley ***(47)***, leading to its map-based isolation ***(38)***.
9. Closely related plant species whose ancestors diverged <10 mya may have retrotransposons with significant identity in addition to genes.
10. The gene prediction programs often annotate retrotransposons and other transposable elements as genes. These elements may also show homology to ESTs, since some of them are transcribed. Therefore, it is important to identify these elements, which contribute to false gene predictions.
11. Polyploid genomes have more than one copy for most genes. Also, many proteins are encoded by multigene families or dispersed duplicated genes. Single gene knock-outs may not show any phenotype or desired change due to the compensation of the function by the duplicated copy of the gene. In these cases, crosses between two homozygous mutant lines with individual knock-out mutations of duplicated copies of the gene would yield, in subsequent generations, a mutant line that is homozygous for inactivation of both genes.

References

1. Flavell, R. B., Bennett, M. D., Smith, J. B., and Smith, D. B. (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **12,** 257–269.
2. Bennett, M. D. (1998) Plant genome values: how much do we know? *Proc. Natl. Acad. Sci. USA* **95,** 2011–2016.
3. Moore, G., Devos, K. M., Wang, Z., and Gale, M. D. (1995) Cereal genome evolution—grasses, line up and form a circle. *Curr. Biol.* **5,** 737–739.
4. Gale, M. D. and Devos, K. M. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA* **95,** 1971–1974.
5. Paterson, A. H., Lin, Y. R., Li, Z. K., et al. (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* **269,** 1714–1718.
6. Devos, K. M., Atkinson, M. D., Chinoy, C. N., et al. (1993) Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor. Appl. Genet.* **85,** 673–680.
7. Zhang, H., Jia, J., Gale, M. D., and Devos, K. M. (1998) Relationship between the chromosomes of *Aegilops umbellulata* and wheat. *Theor. Appl. Genet.* **96,** 69–75.
8. Tanksley, S. D., Ganal, M. W., Prince, J. P., et al. (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132,** 1141–1160.
9. Boutin, S. R., Young, N. D., Olson, T., Yu, Z.-H., Shoemaker, R. C., and Vallejos, C. (1995) Genome conservation among three legume genera detected with DNA markers. *Genome* **38,** 928–937.
10. Lagercrantz, U., Putterill, J., Coupland, G., and Lydiate, D. (1996) Comparative mapping in *Arabidopsis* and *Brassica*: fine scale genome colinearity and congruence of genes controlling flowering time. *Plant J.* **9,** 13–20.
11. Livingstone, K. D., Lackney, V. K., Blauth, J. R., van Wijk, R., and Jahn, M. K. (1999) Genome mapping in *Capsicum* and the evolution of genome structure in the Solanaceae. *Genetics* **152,** 1183–1202.
12. Bennetzen, J. L. (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12,** 1021–1030.
13. Bennetzen, J. L. and Ramakrishna, W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol. Biol.* **48,** 821–827.
14. Chen, M., SanMiguel, P., Oliveira, A. C., et al. (1997) Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. *Proc. Natl. Acad. Sci. USA* **94,** 3431–3435.
15. Feuillet, C. and Keller, B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl. Acad. Sci. USA* **96,** 8265–8270.
16. Tikhonov, A. P., SanMiguel, P. J., Nakajima, Y., Gorenstein, N. M., Bennetzen, J. L., and Avramova, Z. (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA* **96,** 7409–7414.
17. Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000) The

complete sequence of 340 kb of DNA around the rice *Adh1–Adh2* region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* **12,** 381–391.
18. Dubcovsky, J., Ramakrishna, W., SanMiguel, P., et al. (2001) Comparative sequence analysis of colinear barley and rice BACs. *Plant Physiol.* **125,** 1342–1353.
19. Ramakrishna, W., Dubcovsky, J., Park, Y.-J., et al. (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162,** 1389–1400.
20. Ku, H.-M., Vision, T., Liu, J., and Tanksley, S. D. (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97,** 9121–9126.
21. Rossberg, M., Theres, K., Acarkan, A., et al. (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. *Plant Cell* **13,** 979–988.
22. Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G., and Tanksley, S. (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* **14,** 1441–1456.
23. Lagercrantz, U. (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* **150,** 1217–1228.
24. Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000) Extensive duplication and reshuffling in the *Arabidopsis thaliana* genome. *Plant Cell* **12,** 1093–1101.
25. Grant, D., Cregan, P., and Shoemaker, R. C. (2000) Genome organization in dicots. I. Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **97,** 4168–4173.
26. O'Neill, C. and Bancroft, I. (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var *alboglabra* that are homoeologous to sequenced regions of the chromosomes 4 and 5 of *Arabidopsis thaliana*. *Plant J.* **23,** 233–243.
27. Allen, K. D. (2002) Assaying gene content in *Arabidopsis. Proc. Natl. Acad. Sci. USA* **99,** 9568–9572.
28. Laurie, D. A. and Devos, K. M. (2002) Trends in comparative genetics and their potential impacts on wheat and barley research. *Plant Mol. Biol.* **48,** 729–740.
29. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., and Bennetzen, J. L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20,** 43–45.
30. SanMiguel, P. J., Ramakrishna, W., Bennetzen, J. L., Busso, C. S., and Dubcovsky, J. (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome $5A^m$. *Funct. Integr. Genomics* **2,** 70–80.
31. Avramova, Z., Tikhonov, A., SanMiguel, P., et al. (1996) Gene identification in a complex chromosomal continuum by local genomic cross-referencing. *Plant J.* **10,** 1163–1168.

32. Bennetzen, J. L. and Ramakrishna, W. (2002) Exceptional haplotype variation in maize. *Proc. Natl. Acad. Sci. USA* **99,** 9093–9095.
33. Kilian, A., Chen, J., Han, F., Steffenson, B., and Kleinhofs, A. (1997) Towards map-based cloning of the barley stem rust resistance genes *Rpg1* and *rpg4* using rice as an intergenomic cloning vehicle. *Plant Mol. Biol.* **35,** 187–195.
34. Leister, D. M., Kurth, J., Laurie, D. A., et al. (1998) Rapid reorganisation of resistance gene homologues in cereal genomes. *Proc. Natl. Acad. Sci. USA* **95,** 370–375.
35. Pan, Q. L., Liu, Y.S., Budai-Hadrian, O., et al. (2000) Comparative genetics of nucleotide binding site leucine-rich repeat resistance gene homologues in the genomes of two dicotyledons: tomato and *Arabidopsis*. *Genetics* **155,** 309–322.
36. Arumuganathan, K. and Earle, E. D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Reporter* **9,** 211–215.
37. Han, F., Kilian, A., Chen, J. P., et al. (1999) Sequence analysis of a rice BAC covering the syntenous barley *Rpg1* region. *Genome* **42,** 1071–1076.
38. Brueggeman, R., Rostoks, N., Kudrna, D., et al. (2002) The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. *Proc. Natl. Acad. Sci. USA* **99,** 9328–9333.
39. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res.* **8,** 186–194.
40. Gordon, D., Abajian, C., and Green, P. (1998) CONSED: a graphical tool for sequencing finishing. *Genome Res.* **8,** 195–202.
41. Sonnhammer, E. L. L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167,** 1–10.
42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
43. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
44. Devos, K. M., Brown, J. K. M., and Bennetzen, J. L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12,** 1075–1079.
45. Sharp, P. A. (1999) RNAi and double-strand RNA. *Genes Dev.* **13,** 139–141.
46. Colbert, T., Till, B. J., Tompa, R., et al. (2001) High-throughput screening for induced point mutations. *Plant Physiol.* **126,** 480–484.
47. Druka, A., Kudrna, D., Han, F., et al. (2000) Physical mapping of the barley stem rust resistance gene *rpg4*. *Mol. Gen. Genet.* **264,** 283–290.

8

Using Natural Allelic Diversity to Evaluate Gene Function

Sherry R. Whitt and Edward S. Buckler, IV

Summary

Genomics has developed a wide range of tools to identify genes that play roles in specific pathways. However, relating individual genes and alleles to agronomic traits is still quite challenging. We describe how association analysis can be used to relate natural variation at candidate genes with agronomic phenotypes. Association approaches in plants can provide very high resolution and can evaluate a wide range of alleles rapidly. We discuss issues related to experimental design, germplasm sample, molecular assay, population structure, and statistical analysis necessary for association analysis in plants.

Key Words

association analysis, candidate gene, linkage disequilibrium, LD, maize, phenotypic variation, population structure, mapping, QTL, quantitative trait loci, selection, diverse germplasm

1. Introduction

We describe a methodology for dissecting complex traits using association analysis and natural diversity. In a high diversity species such as maize, association analysis has the potential to map quantitative trait loci (QTL) with up to 5000 times better resolution than mapping with standard F_2 populations ***(1)***. In addition, association approaches may survey tens of alleles, whereas standard mapping approaches survey a maximum of two alleles. Association approaches do not require special mapping populations, but rely on the extensive history of mutation and recombination to dissect a trait. The structure of linkage disequilibrium (LD), which is the correlation between polymorphisms, and evaluation of selection is key to utilizing association analysis ***(2,3)***.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

The use of extant natural diversity provides advantages in resolution and breadth of survey, but can also present added difficulties in accurately assessing the true cause of an association. The most serious false positives can result when unlinked markers produce a positive association because of underlying population structure. The complex breeding history of most crops and the limited gene flow in most wild plants creates population stratification within the germplasm ***(4)***.

In recent years, a few statistical methods have been developed that use independent marker loci as a means of detecting and correcting for population structure ***(5,6).*** These methods work on the assumption that population structure should affect all loci in a similar manner. Reich and Goldstein ***(6)*** propose scoring a moderate number of unlinked genetic markers (e.g., single nucleotide polymorphisms [SNPs] or simple sequence repeats [SSRs]) and then comparing the strength of the candidate gene association with those of the unlinked markers. We have utilized a modified approach designed by Pritchard et al. ***(7,8)***, which incorporates a test statistic of likelihood ratios that includes estimates of subpopulation allele frequencies and evaluates quantitative traits ***(1)***.

A standard procedure for carrying out association analysis on candidate genes is as follows (*see* **Fig. 1**):

1. Select positional candidate genes using existing QTL and positional cloning studies.
2. Choose germplasm that will capture the bulk of diversity present. When possible, inbred lines should be used.
3. Score phenotypic traits in replicated trials.
4. Amplify and sequence candidate genes.
5. Manipulate sequence into valid alignments and identify polymorphisms.
6. Obtain diversity estimates and evaluate patterns of selection.
7. Statistically evaluate associations between genotypes and phenotypes taking population structure into account.

2. Materials

2.1. Germplasm

Sample at least 100 inbred lines of germplasm (for a maize example, see [http://www.maizegenetics.net]). For high resolution do not choose closely related samples. In order to test for selection in a crop, collect one sample from a sister taxon to function as an outgroup for the Hudson, Kreitman, and Aguade (HKA) test ***(9)***.

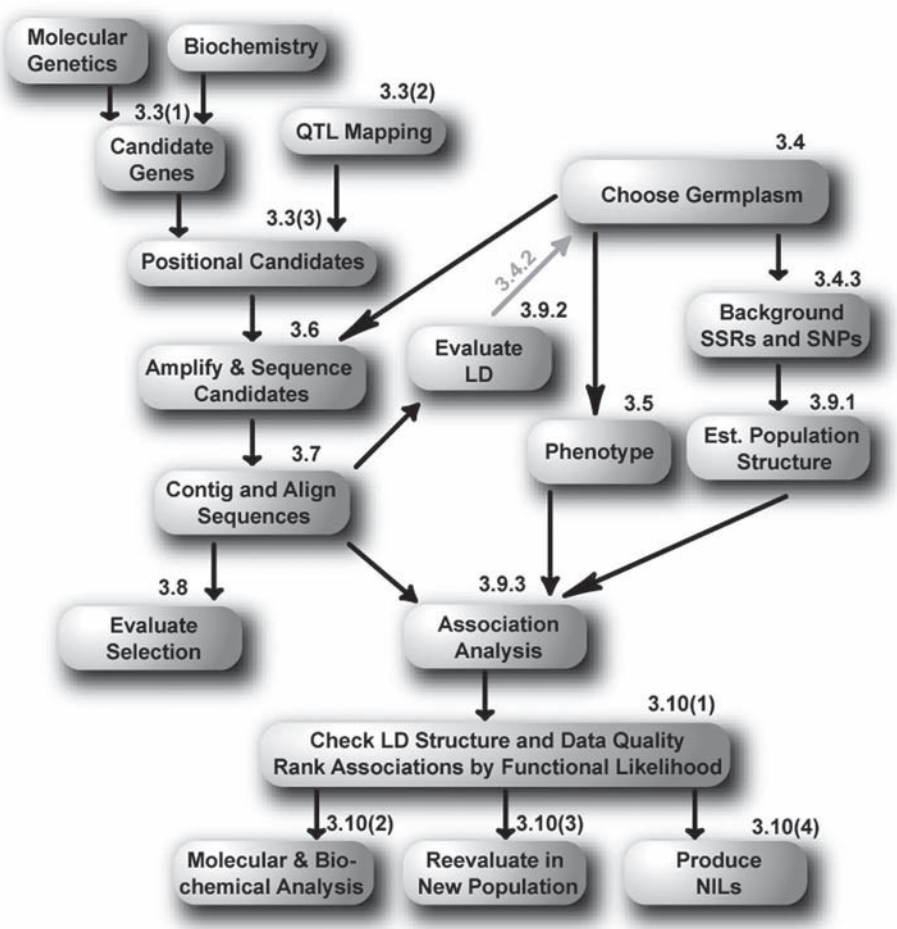


Fig. 1. Association study employs techniques from molecular biology, field sampling–breeding, bioinformatics, and statistics. The steps necessary to associate a particular genotype with a phenotypic trait are illustrated. Above each step is a numeric reference to the relevant text. The gray arrow linking "Evaluate LD" and "Choose Germplasm" signifies the potential need to revise the choice of germplasm once the structure of LD is known.

2.2. Primer Design

Primer3.0 (http://www-genome.wi.mit.edu) and PCR-Overlap (http://droog.mbt.washington.edu) provide oligonucleotide design for Linux or Unix® operating systems.

2.3. PCR

1. FailSafe™ system (Epicentre): 2× premixtures labeled A–L; contain dinucleotide triphosphates (dNTPs), Tris-based buffering solution, varying amounts of $MgCl_2$, and betaine (*see* **Note 1**).
2. FailSafe enzyme: 2.5 U/μL, a mixture of polymerases with proofreading capabilities (*see* **Note 2**).
3. Genomic DNA (33 ng/μL) (purified with DNeasy™ plant maxi kits [Qiagen]).
4. Primers.
5. QIAquick™ 8 PCR purification kit (Qiagen) (*see* **Note 3**).

2.4. Cycle Sequence

1. BigDye™ chemistry (Applied Biosystems) (we dilute enzyme with dilution buffer for quarter reactions). Dilution buffer (halfTERM dye terminating sequence [Sigma] or Half-Dye™ mixture [Denville Scientific]) (see **Note 4**).
2. High-performance liquid chromatography (HPLC) water.
3. Purified PCR template (*see* **Subheading 3.6.6.**).
4. Primers (*see* **Subheading 3.6.1.**).
5. DyeEx™ terminator removal kit (Qiagen).

2.5. Sequence Manipulation Software

1. PHRED and PHRAP versions from CodonCode (http://www.codoncode.com/) are used to assess sequence quality and contig (join) sequences ***(10)***.
2. Biolign (Tom Hall [http://www.mbio.ncsu.edu/BioEdit/bioedit.html]) is used to edit multiple alignments of contigs and evaluate SNPs. Biolign is a custom software package. MegAlign from DNASTAR and Sequencher™ from GeneCodes (http://www.genecodes.com) offer some similar features.

2.6. Software for Testing for Selection

Statistical analyses for several tests of selection are performed with DnaSP 3.0 (http://www.bio.ub.es/~julio/DnaSP.html), which has a user-friendly interface ***(11)***, and SITES (http://lifesci.rutgers.edu/~heylab) ***(12)***.

2.7. Association Analysis Software

1. Population structure software: STRUCTURE (http://pritch.bsd.uchicago.edu) is an excellent program to estimate population structure ***(7)***.
2. LD software: Arlequin (http://lgb.unige.ch/arlequin) can handle a wide range of markers and sequences ***(13)***. It can also calculate LD from genotypic data. DnaSP (http://www.bio.ub.es/~julio/DnaSP.html) manages numerous DNA sequences and can plot LD ***(11)***. PowerMarker (http://www.powermarker.net) can incorporate a wide range of markers and genotypic data and can produce plots of LD. TASSEL (http://www.maizegenetics.net) has the capability to cope with a wide range of markers, sequences, and plot LD.
3. Association software: SAS (http://www.sas.com) is a general purpose statistical

software package and can carry out a wide range of statistics useful for association analysis. STRAT (http://pritch.bsd.uchicago.edu) can be used for testing association of binary traits across structured populations ***(8)***. TASSEL (http://www.maizegenetics.net) can perform analysis of variance (ANOVA) and logistic regression association tests that control for population structure.

3. Methods

3.1. Polymerase Chain Reaction

1. Combine the following, scaling vol for number of reactions desired, to produce 25 μL total vol reactions (add genomic DNA subsequently): 12.5 μl 2× premixture, 7.0–9.0 μL HPLC-grade water, 1.0 μL-20 μ*M* forward primer, 1.0 μL 20 μ*M* reverse primer, 0.5–1.0 μL *Taq* DNA polymerase, and 1.0–3.0 μL genomic DNA.
2. Amplify the above reaction in a thermal cycler.
3. Purify PCR product using QIAquick 8 PCR purification kit (*see* **Note 3**).

3.2. Cycle Sequence

1. Prepare a standard reaction as follows in 10 μL total vol: 2 μL terminator enzyme, 2 μL dilution buffer, 4 μL PCR template, and 2 μL 3–5 μM primer.
2. Sequence reactions are cleaned via DyeEx terminator removal kit (for cycle profile *see* **Subheading 3.6.6.**).

3.3. Selection of Positional Candidate Genes

Choosing candidate genes, i.e., those genes most likely to contain the polymorphism responsible for the phenotype, is one of the most critical steps in conducting association analysis. However, candidate gene selection is currently as much an art as a science. We refer to candidate genes that fall within QTL intervals as positional candidate genes. QTL mapping often has limited resolution, but is an excellent way to narrow the search for candidates to specific chromosomal regions. Focusing on positional candidate genes will maximize the opportunity to find associations. The major aspects of choosing genes are:

1. Collect a list of genes that affect the phenotype of interest. Mutagenesis, biochemistry, various profiling technologies, comparative genomics, and positional cloning techniques–studies can aid in the identification of genes. Create a list of chromosomal positions for these candidate genes.
2. Collect a list of map positions of QTL for the trait of interest over all previous experiments. Various databases such as MaizeDB (http://www.agron.missouri.edu) and Gramene (http://www.gramene.org) provide a good starting point for determining these positions. A candidate gene should not be ruled out if only one or two populations have been mapped.
3. Compare the two lists and generate a list of all known genes with potential phe-

notypic effects in QTL confidence intervals. These positional candidate genes with the most neighboring QTL are most likely to have segregating variation at the locus. These positional candidate genes should be sampled first.

3.4. Choice of Germplasm

3.4.1. Phenotypic Diversity

The choice of germplasm is crucial to the discovery of useful alleles. In order to have enough statistical power to find an association, it is critical that the samples span the full range of phenotypic variation. To maximize the range of alleles tested, a genotypically diverse set of germplasm should be chosen. When available, marker or phenotypic surveys can be used to choose a subset of the germplasm that is most diverse. Software such as MSTRAT (http://www.ensam.inra.fr/gap/MSTRAT/mstratno.htm) and PowerMarker provide methods for helping to choose the germplasm. From a practical level, we found that a sample of 100 diverse inbred lines has enough statistical power to identify associations that control 10% of the phenotypic variation ***(1)***. Larger samples and/or more replications of phenotypic evaluation could be used to identify associations with smaller effects. Although geneticists are forced to do association studies with outbred populations, inbred lines provide a number of advantages to plant researchers. Inbred samples allow direct identification of haplotypes throughout the genome, generally have more consistent phenotypes than segregating populations, and provide evaluation of phenotype without the complications of dominance. In some species, core sets of germplasm have been defined and characterized, and these are excellent starting points for association studies.

3.4.2. Resolution and LD

The choice of germplasm will also determine the resolution of association approaches. Highly diverse germplasm has an extensive history of recombination, which can result in high-resolution association analysis. However, high resolution will require a high marker density to identify associations. Resolution of associations is directly related to the structure of LD ***(14)***. LD is the correlation between pairs of polymorphisms. One simple way to estimate LD between pairs of sites is to calculate r^2 ***(15)***. The average distance between polymorphisms, at which r^2 drops below 0.1, is a rough estimate of the resolution within a specific population. The rate of LD decay in most cases needs to be determined empirically for any given population ***(2,16)*** (*see* **Note 5**). However, the rate of LD decay may also be locus-specific, as differences in recombination rate, mutation rate, and selection history can affect LD patterns.

3.4.3. Germplasm Population Structure

The final consideration in selecting the sample population is whether to use randomly or nonrandomly mated germplasm. Unfortunately, there is little truly randomly mated breeding germplasm available, other than a few unselected synthetic populations. Many of these randomly mated populations represent a rather narrow group of germplasm, which is likely to lower resolution and harbor only a narrow range of alleles. However, if nonrandomly mated germplasm is used, population structure needs to be controlled in the statistical analyses. In addition, the genome for each sample population should be genotyped with SSRs, SNPs, restriction fragment-length polymorphisms (RFLPs), random-amplified polymorphic DNAs (RAPDs), or amplified fragment-length polymorphisms (AFLPs) to provide an estimate of population structure (*see* **Subheading 3.9.1.**). Our experience has been that 50–150 markers generally provide good estimates of population structure. The ideal markers are either a modest number of SSRs or large numbers of SNPs, while if resources are limited, AFLP may provide a good compromise.

3.5. Phenotypes

We test for associations between polymorphisms with a wide range of agronomic and physiological phenotypic traits. The phenotypic data comes from 2 to 3 field seasons of randomized plots with 10–15 plants per row, replicated across multiple environments. The measurement of phenotypic traits needs to be a balance of simplicity in data collection, biological relevancy, and reproducibility.

3.6. Gene Amplification and DNA Sequencing

Once a candidate gene has been identified, the researcher carries out a set of standard procedures including various molecular techniques. A general guideline is as follows:

1. Design compatible primer pairs from candidate gene sequence.
2. Employ PCR to amplify the target.
3. Verify product from PCR by agarose gel electrophoresis and purify the DNA.
4. Obtain a DNA sequence product directly from the PCR product, by using commercially available labeling chemistry and enzymes (*see* **Subheading 3.6.5.**).
5. Clean up sequencing reactions to eliminate excess dNTPs, enzyme, and buffer.
6. Determine nucleotide sequence by electrophoresis (*see* **Note 6**).

3.6.1. Primer Design

1. Define a “standard” allelic sequence for primer design and future alignments (*see* **Subheading 3.7.2.**).

2. Design a series of overlapping primers, based on the standard allelic sequence, across the gene via PCR-overlap in conjunction with Primer3 (*see* **Note 7**). Typical coverage is usually in 1-kbp fragments of the gene. Primers may also be designed manually by visual inspection of sequence and apply the general rules of primer design. Generate forward and reverse primers approx 18–25 bp in length with similar melting temperature (T_m) (near 60°C) and order from one of several companies offering this service (*see* **Note 8**).
3. Resuspend lyophilized forward and reverse oligonucleotides at a standard stock concentration of 100 μ*M* in 1× Tris-ethylene diamine tetraacetic acid (EDTA) buffer and store at –80°C. Oligonucleotides should be further diluted to a working stock concentration of 20 μ*M* for PCR and 3–5 μ*M* for cycle sequencing, both of which should be stored at –20°C.

3.6.2. Optimization of PCR

1. Attempt initial PCRs by combining various primer sets and buffer conditions at an annealing temperature gradient from 50°–60°C. Utilize genomic DNA from a few representative test samples before including the entire population. Buffers containing a range of $MgCl_2$ and betaine are evaluated for optimal amplification (*see* **Note 9**).
2. A standard PCR program carried out on a thermal cycler may include: 5 min denaturation at 96°C, followed by 25–35 cycles of: 30 s denaturation at 96°C, 30 s annealing at 50°–65°C, and 30 s to 4 min extension at 70°–72°C; a final extension at 70°–72°C for 5–10 min, and hold at 4°C. A typical PCR program will take from 2–4 h (*see* **Note 10**). Annealing temperatures are set a couple of degrees below the primer melting temperatures, and extension times are delineated by the size of the expected PCR product using the 1 min/1 kbp rule.
3. When optimal buffering conditions and annealing temperatures are found, the remainder of the sample population is included along with numerous negative and positive controls in subsequent PCR (*see* **Note 11**). Increased product amount can be obtained by scaling up the total reaction vol to 50 μL.
4. Most PCR products can be directly sequenced from inbred lines, because all loci are homozygous. Some researchers may wish to compare sequence diversity between domesticated species and wild relatives (e.g., *Zea mays* ssp. *mays* and *Zea mays* ssp. *parviglumis*). Due to the heterozygous nature of wild relatives, we clone the PCR product before sequencing.

3.6.3. Agarose Gel Electrophoresis

Once PCR is completed, check the product for the correct band size and amount by agarose gel electrophoresis with ethidium bromide staining. Utilize a mass ladder as a standard to determine size and quantity of product fragments. Add 6× loading dye to the mass ladder and samples prior to electrophoresing the samples at an appropriate voltage. Visualize the products by UV transillumination (*see* **Note 12**). When more than one product is obtained, the correct fragment may be excised from the agarose gel with a sterile blade.

3.6.4. Purification of PCR Product

PCR product will yield quality sequence data only when all enzyme, primer, and other reagents are removed from the reaction. Final products are purified with a vacuum manifold in an 8-strip format. The procedure takes approx 20 min for 48 samples. Product is eluted into a 96-well plate and maintained at –20°C until the template is sequenced.

3.6.5. Cloning Gene Fragments

TOPO® TA cloning kit (Invitrogen) provides superior success rates. The PCR product is ligated into an approx 4-kbp vector with thymidine overhangs and transformed into chemically competent *Escherichia coli* One Shot® TOP10® (which provides both ampicillin and kanamycin resistance, as well as blue–white colony screening). To prepare small quantities of DNA, we use Qiagen's QIAprep® 8 turbo miniprep kit. TOPO TA cloning and transformation takes approx 2 h and overnight growth in an incubator at 37°C. The miniprep preparation by vacuum manifold takes 30 min to complete. Ultimately, clones are stored as glycerol stocks in 96-well format at –80°C.

3.6.6. Cycle Sequencing

Sequence reactions should follow standard protocols for the chemistry and, in general, take approx 2 1/2 h to complete.

1. The sequence reaction is set up in a 10-µL total vol (*see* **Subheading 3.2.**).
2. The amount of template used should equate to approx 30–50 ng of DNA.
3. A typical cycle sequence program is: 92°C for 30 s, 50°C for 30 s, and 60°C for 4 min, repeated 25 times, then maintain at 4°C until removal from thermal cycler.
4. Sequence product is cleaned up to eliminate excess enzyme and primer. Prepackaged kits are available with a procedure that takes approx 20 min (*see* **Subheading 3.2.**).
5. DNA Sequence is obtained by Model 3700 capillary electrophoresis (Applied Biosystems) in a 96-well format. Sequence products are light-sensitive. Keep exposure to a minimum.

3.7. Sequence Manipulation

Sequence manipulation involves database handling of trace files, applying quality scores to individual bases, and contiging (joining) and aligning sequence data.

3.7.1. Join Sequence Fragments

DNA sequence is received as trace files of the chromatograms and text files of the nucleotide sequence. The trace files are sorted by gene and sample to

individual gene folders, accordingly. Quality scores and contigs are obtained using the CodonCode versions of PHRED and PHRAP (*see* **Note 13**). The quality scores are reported in spreadsheet format as the total number of bases with a Phred score of 20 or higher for a particular sample. Sequences with more than 400 bases with scores of 20 and higher are included in the alignment. CROSSMATCH is used to remove vector sequence if fragments were cloned. Phrap contigs the sequences to produce an ".ace" file, which contains nucleotide reads and associated Phred scores.

3.7.2. Align Sequences

.ace files are aligned in the software program Biolign. When possible, a published sequence is used as the standard in a framework around which the alignment is built. The standard sequence can be used to delineate base call differences in any of the sequenced samples. Phrap quality coloring indicates the quality of sequence, by highlighting specific bases with different colors based on phred quality scores <30. Regions with low quality (e.g., <Phred 20) are converted to the missing data symbol "?" or "N." Prior to calculating diversity indices, introns and exons are delimited, and then the annotated contiged sequence is saved in the NEXUS file format.

3.8. Evaluation of Diversity and Selection

The power to detect associations depends on genotyping, genetic architecture, and accurate phenotypic evaluations. If there are complications with any of these three factors, there may be little statistical power in relating a gene to a specific phenotype. However, the signature of artificial selection can also be used to provide evidence that a specific gene is important for controlling phenotypic variation. If a gene has been a target of selection through the domestication and breeding process, then it is likely to control an agronomic phenotype and could be useful in future breeding and genetic manipulation. Nucleotide diversity surveys can powerfully detect several forms of selection. We describe two tests of selection that can be useful in finding genes that play key roles in phenotypic variation. The Tajima's D test evaluates diversity within a species to find evidence of selection, while the HKA test compares nucleotide diversity within a species to the nucleotide difference with a related species. Diversity and selection estimates can be obtained as follows:

1. DnaSP enables the user to calculate several indices of genetic diversity, divergence, and selection. Generate an aligned nucleotide sequence with codon assignment in NEXUS file format (*see* **Note 14**).
2. Calculate diversity measures for the sequence data. We report π (the average number of nucleotide differences per site between two sequences) and θ (similar

to π, but focuses on the number of segregating sites) for nonsynonymous and synonymous sites and LD (*see* **Subheading 3.9.2.**).

3. Perform tests of selection. Tajima's D test statistic compares diversity based on average number of differences (π) vs the number of segregating sites (θ), hypothesizing all mutations are selectively neutral ***(17)***. The statistic may also reflect demographic changes or population structure, so caution is needed in interpreting these results.
4. The HKA test examines the ratio of intraspecific diversity to interspecific divergence using an outgroup ***(9)*****.** The outgroup species should have diverged just before the time when the alleles within the target species began diverging (*see* **Note 15**). Within the DnaSP program, a second data window must be opened containing the outgroup sequence. The test is calculated by comparing silent θ and silent K (divergence). A low value relative to other loci suggests that selection, specifically, has reduced diversity at a particular locus. Neutral loci are needed for comparison in this test.
5. A significant selection test may mean little molecular variation with which to find associations. These tests indicate that selection has occurred, but they are generally ambiguous as to why selection has occurred.

3.9. Statistical Applications to Find Genotype–Phenotype Associations

3.9.1. Estimate Population Structure

If the samples are not randomly mated, it is critical that population structure be included in the association analysis. The STRUCTURE software is a good way to estimate population structure for association approaches.

1. Convert genotypic marker data (e.g., random SSR or SNP data throughout genome) to STRUCTURE format (*see* **Note 16**).
2. Run STRUCTURE and test with one population, continue to increase the population number until the maximum likelihood is identified. Cycles (100,000) for both burn-in (the period where the model explores the parameter space) and likelihood estimation seems to work well. At least five repetitions should be conducted for each population size (*see* **Note 17**).
3. Extract the Q matrix from the optimal result for later use (*see* **Subheading 3.9.3.**).

3.9.2. Evaluate LD

Understanding the structure of LD for a specific locus will, in turn, reveal the association resolution possible at that locus. For example, if LD decays within 1000 bp, then 1 or 2 markers per 1000 bp will be needed to identify associations.

1. DnaSP, Arlequin, or TASSEL will calculate LD between pairs of polymorphisms (r^2 or D'). Use any one of these programs to calculate all pairwise estimates of LD.

2. Plot the distance between the polymorphisms in basepairs vs LD (e.g., r^2). From this plot, one can estimate the point at which r^2 is below 0.1 (a rough estimate of the resolution of the association study).
3. Plot the strength of LD between all pairs of sites, which can be graphically done in PowerMarker or TASSEL. The graph will identify blocks of high LD and will show which sets of sites are highly correlated. Association approaches will have trouble differentiating between blocks of highly correlated sites.

3.9.3. Evaluate Associations

1. Filter polymorphisms: the segregating sites need to be extracted from the sequence alignments either by hand or by programs such as TASSEL and DnaSP. Normally, polymorphisms that are present in less than three samples or with a frequency <5% are not included in the analyses. These low frequency polymorphisms may be the product of PCR or sequencing error. Additionally, there is rarely enough statistical power to test for association at these low frequency polymorphisms. Insertions and deletions also need to be identified and coded for analysis. TASSEL does this automatically, while DnaSP ignores this type of polymorphism.
2. Randomly mated samples: when samples are truly randomly mated, no correction for population structure is required. If the trait is binary (e.g., yellow vs white kernels), then a series of chi-square tests (χ^2) can be used to evaluate whether the segregating polymorphisms associate. If the trait is quantitative, then a series of *t*-tests or ANOVA can be used to evaluate the associations (*see* **Note 18**).
3. Structured samples: when population structure is present, statistical analysis must account for it. If the trait is binary, the STRAT program can be used to evaluate the associations. If the trait is quantitative, either SAS or TASSEL can be used to implement the logistic regression ratio test. In the null hypothesis H_0, candidate polymorphisms are independent of phenotype; while in the alternative hypothesis H_1, candidate polymorphisms are associated with the phenotype. The probability of each hypothesis is compared in the following way:

$$\Lambda = \frac{\Pr_1(C;T,\hat{Q})}{Pr_0(C;\hat{Q})}$$

Where *C* is the genotype of the candidate polymorphism for all lines, and *T* is the trait value for all lines. In this test, the difference in the natural logarithm likelihoods of the model with (Pr_1) and without (Pr_0) the trait is the test statistic Λ (*see* **Note 19**). Since the distribution of Λ is not known precisely, permutations should be used to determine significance. If several sites with high LD are being scored, then the maximum Λ over all sites is used as the test statistic Λ_{max}. Permutations are calculated based on this Λ_{max} statistic.

4. Permutations to determine significance: these statistical tests will result in a *P*-value associated with each polymorphism-trait pair. However, for many associa-

tion tests, there will be 10s or 100s of polymorphisms to test. The normal modification for multiple tests would be a Bonferroni correction, however, this is far too conservative for highly correlated polymorphisms. The goal of permutations is to determine the number of independent tests, which is confounded by LD, and account for nonnormality in trait distributions. The trait values should be permuted relative to the fixed haplotypes ***(18)***, and then associations recalculated for 100–1000 permutations. Pritchard et al. ***(8)*** suggests permutations based on population structure, and this approach is implemented in STRAT and TASSEL.

5. Compare the permuted *P*-value to the distribution of *P*-values for random markers across the genome. In some cases, the estimates of population structure do not explain all of the structure. Subsequently, the random markers used for estimating population structure could be used, as could data from unrelated candidate genes. The candidate gene *P*-value could be rescaled based on the *P*-values for the random markers. For example, if the candidate gene had a *P*-value of 0.03, but 7% of the random markers had a *P*-value <0.03, then the candidate genes *P*-value could be rescaled to 0.07. This is probably a conservative test, as some of the random markers are likely to be truly associated with the trait.

3.9.4. Evaluate Associations Using TASSEL

Outlined below is a step-by-step example of how to use TASSEL to carry out association tests with structured populations. TASSEL can work with data stored in databases, but in this example, we describe TASSEL use with flat files (unclick the DB button in the main window).

1. Download and install the program by going to (http://www.maizegenetics.net).
2. Create a sequence alignment in PHYLIP format or CLUSTAL format. Many sequence editors can produce these alignments, such as CLUSTALW or BioEdit. Load the sequence alignment into TASSEL by clicking the Data button and then the Gene button and selecting your alignment file.
3. Create text files in the format described in the TASSEL help section for the population structure matrix (Q matrix from 3.7.1) and the trait data. Load the trait file by clicking on the Trait button, and the Q matrix by clicking the Pop button. It is critical that taxa names are exactly the same for the sequence alignment, Q matrix, and trait matrix.
4. Remove the invariant and low frequency sites from the sequence alignment by selecting the sequence alignment and clicking the Sites button. We normally examine sites with a minimum frequency of 0.05, as less frequent sites often have little power to detect significant results with samples less than several hundred taxa.
5. Join the filtered alignment with the Q and trait matrices by selecting all three matrices and the clicking the ∩ Join button, which will produce the intersection of these datasets.
6. Click Analysis and then Struct. Assoc. to carry out a structured association analy-

sis. Use the arrows to move the Q matrix values to the Pop Structure Estimate list. Generally at least 1000 permutations should be run.

7. These results will be summarized in two reports. The first report summarizes the results for the entire data set and accounts for the multiple tests conducted. The second report provides information on how individual sites were associated. Results may be viewed in tabular or graphical format by clicking the Results button.

3.10. Interpretation of Genotype–Phenotype Associations

Once an association is empirically determined, the validity of the association must be ascertained.

1. Which associating polymorphisms most likely control the trait? First, it is critical that genotypes be rechecked, and results should be examined to determine if phenotypic outliers are driving associations. Association studies will often find multiple polymorphisms that significantly associate. Carefully examining the LD structure surrounding the association can help identify this suite of polymorphisms and where more sampling may be needed. Although the most significant site is the most likely cause for the association, many of the slightly less significant sites could actually be the functional cause of the phenotypic variation. We find that breaking the polymorphisms into likely functional (biologically significant) vs likely silent is useful in developing lists of sites for future evaluation. Radical coding sequence changes, changes in conserved promoter motifs, changes within splicing motifs, and large insertions–deletions are generally put in the likely functional list.
2. The most straightforward way to prove an association is to evaluate the candidate polymorphisms in an entirely different population sample. Only polymorphisms that are closely linked to the cause of a phenotype should be significant in a second study. It is important that the population structure of the second sample is truly independent of the first sample. Only the candidate polymorphisms need to be retested in the new sample. In maize, we are using randomly mated synthetic populations for reevaluation of association studies.
3. In some cases, associations will suggest a molecular or biochemical mechanism of action. Following-up hypotheses generated by association analysis with molecular biology and biochemistry can be very productive, but it should be warned that association studies could be picking up on effects that only explain a few percent of the variation. Many biochemical and molecular approaches may not be quantitatively sensitive enough to detect such small changes at a molecular level.
4. Final proof of the association can be obtained through marker-assisted selection and production of near isogenic lines (NIL).

3.11. Conclusions

Mapping with F_2 or derived populations is powerful for evaluating two alleles with low resolution. In contrast, association analysis can evaluate numerous alleles at high resolution. These two approaches are complementary. The successful integration of these two approaches will allow the rapid dissection of almost any trait within a few years time. The key to association analysis is the choice of germplasm, quality of phenotypic data, and use of statistical analyses to control for population structure. The combination of association mapping and QTL mapping could make it routine to dissect complex traits down to the single gene level.

4. Notes

1. Maize is particularly GC-rich and requires additional components for optimal PCR; the FailSafe system is expensive, but allows for easier optimization as necessary reagents are premixed and contain a wide range of concentrations.
2. We have had success with *Taq* DNA polymerase (no proofreading) as well.
3. These kits provide exceptional product for subsequent cycle sequencing, but are fairly expensive. Phenol–chloroform extraction is also used as an inexpensive alternative.
4. We tried a homemade recipe, but achieved 100–200 fewer bases with sequence results. Others have had success with a solution containing Tris, $MgCl_2$ at pH 9.0, and water.
5. In maize for example, LD decays within 600 bp for landraces of maize ***(16)***, within 2000 bp for diverse breeding inbred lines ***(2)***, whereas LD persists up to 100,000 bp for elite inbred lines ***(14)***.
6. We typically obtain 500–600 bp reads on average from a Model 3700 analyzer. Alternate sequence methodologies are available, such as Model 377 and MegaBase technologies.
7. We include a library of common repetitive elements in the mispriming library, which seems to improve efficiency especially for longer amplicons.
8. We use Operon Technologies, Illumina, and Oligos, etc., for primer synthesis. Primer is delivered at room temperature in pellet form. We order at the 50 nmol scale with no additional purification.
9. Epicentre technologies does not provide specific information on reagent concentrations for the Failsafe 2X premixtures. The "midrange" refers to premixtures "D," "E," "F," and "G."
10. Often a PCR thermal cycler program utilizing a two-step or touchdown method is superior. This methodology allows increased specificity of primer annealing by carrying out the first 10 cycles at a fairly high annealing temperature (e.g., 60°–65°C) and the remaining cycles at a temperature approx 7°–10°C lower, aimed at boosting the yield (e.g., 50°–58°C).
11. Optimization of PCR largely depends upon the gene under investigation, espe-

cially GC content and inherent diversity. Ultimately, we have found that the more difficult a particular inbred line is to amplify, the more interesting the nucleotide sequence. Quite often, certain inbred lines require separate optimization to obtain PCR product, and even then, sequencing can require "line-specific" primers where highly polymorphic regions exist. We strive to obtain 2× coverage over the entire gene for almost all the samples.

12. We gauge the bandwidth to determine appropriate elution vol on purification by rough visual quantification. Mainly, we check to ensure there is a single band of the expected size present. Typically, we use 30 ng DNA/reaction for sequencing. Even less concentrated product may yield adequate sequence results. Only very weak bands, as visualized by gel electrophoresis, will yield poor results (e.g., <15 ng DNA/8 μL PCR product).
13. PHRED and PHRAP allow base calling and assembly of DNA sequence by simple Fourier methods.
14. Gaps are treated as missing data, and all sites at those positions are excluded from analyses; gaps in exons may alter the translation.
15. For example, in maize we use *Tripsacum*, which diverged from maize about 5 million yr ago, while the allelic diversity in maize is roughly 1 to 2 million yr old.
16. If inbred lines are being used, we set the second allele to missing (–9) as it eliminates the Hardy-Weinberg part of the model, and helps reconstruct the population structure before the inbreeding.
17. Sometimes the model seems to split off individual taxa, however, these single taxa populations are not very useful for controlling population structure. The user may want to try the Q matrix based on the population number before the individual taxa populations are split off.
18. The described approach analyzes individual sites, however, the analysis of haplotypes can also be powerful statistically. There are many approaches in the human genetic literature that could be used.
19. The SAS script for the test is as follows:

```
proc logistic data = indata outest = resultH0;
model testPolymorphism = Q1 Q2; run;
proc logistic data = indata outest = resultH1;
model testPolymorphism = trait Q1 Q2; run;
```

Then the difference of _LNLIKE_ of both tests is used as the test statistic.

Acknowledgments

We thank Jeffry Thornsberry, Brad Rauh, Sandra Andaluz, Sherry Flint-Garcia, Susan Wiltse, and Larissa Wilson for helping to develop these methods and commenting on this manuscript. This research was supported by National Science Foundation (NSF) grant DBI-9872631 and the United States Department of Agriculture–Agricultural Research Service (USDA-ARS).

References

1. Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S., IV. (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* **28,** 286–289.
2. Remington, D. L., Thornsberry, J. M., Matsuoka, Y., et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* **98,** 11479–11484.
3. Nordborg, M., Borevitz, J. O., Bergelson, J., et al. (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30,** 190–193.
4. Sharbel, T. F., Haubold, B., and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9,** 2109–2118.
5. Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65,** 220–228.
6. Reich, D. E. and Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* **20,** 4–16.
7. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959.
8. Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* **67,** 170–181.
9. Hudson, R. R., Kreitman, M., and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159.
10. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185.
11. Rozas, J. and Rozas, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15,** 174–175.
12. Hey, J. and Wakeley, J. (1997) A coalescent estimator of the population recombination rate. *Genetics* **145,** 833–846.
13. Schneider, S., Roessli, D., and Excoffier, L. (2000) *Arlequin ver. 2.000: A Software for Population Genetics Data Analysis*. Genetics and Biometry Laboratory, University of Geneva, Switzerland.
14. Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics and breeding. *Curr. Opin. Plant Biol.* **5,** 94–100.
15. Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38,** 226–231.
16. Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA* **98,** 9161–9166.
17. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.
18. Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.

9

Quantitative Trait Locus Analysis as a Gene Discovery Tool

Michael D. McMullen

Summary

Quantitative trait locus analysis has been a mainstay approach for obtaining a genetic description of complex agronomic traits for plants. What is sometimes overlooked is the role quantitative trait locus (QTL) analysis can play in identifying genes that underlay complex traits. In this chapter, I will describe the basic steps required to conduct QTL analysis in crop plants. This process involves choices by the investigator on type of population to be studied, molecular marker system to be used to genotype the population, and methods for QTL analysis. Examples of cloned genes, first identified as QTL, are also given to persuade the reader of the power of QTL analysis to discover genes controlling traits and phenotypes.

Key Words

QTL, candidate gene, molecular markers, SSR, RFLP, SNP, NIL maize, agronomic traits, genetic mapping, positional cloning

1. Introduction

The past 15 yr have seen an explosion of information on the structure, organization, and gene functions of plant genomes, including the development of high-density molecular marker maps. One application of new mapping technologies has been the genetic dissection of quantitative agronomic traits with much greater precision than was previously possible ***(1)***. Using quantitative trait locus (QTL) analysis, a description of the number, genetic effects, and chromosomal positions for genes underlying many agronomic traits have been obtained. We have clearly entered a new era with the application of genomics to crop improvement. The plant research community now has substantial expressed sequence tag (EST) resources, physical map resources, and access to

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

bioinformatic tools for most major crops, along with full genome sequence of both dicot, *Arabidopsis thaliana*, and monocot, rice (*Oryza sativa* L.) models. Much of the excitement surrounding genomics is the promise that genomics provides for systematic gene discovery, and in using genomics, we can finally reach the "Holy Grail" of cloning and characterizing genes controlling agronomic traits. As recently demonstrated by the cloning and characterization of QTLs from tomato (*Lycopersicon* sp.) ***(2)*** and rice ***(3)***, QTL analysis can be a powerful complementary technology with genomics to "discover" and isolate the genes of most interest to the agronomist. It is the innate ability of QTL analysis to identify those genes that regulate or control variation in phenotypic expression, the raw material of the plant breeder, that requires the integration of QTL and genomic approaches for crop improvement. There are numerous excellent reviews of QTL methodology ***(4–6)***. In this paper, I address some of the considerations and practical choices that face the scientist wishing to use QTL analysis for identifying the genes underlying agronomic traits. I will outline the choices the investigator needs to make in population structure, molecular marker type, and trait analysis necessary to identify QTLs and discuss QTL analysis in the context of gene discovery.

2. Materials

2.1. Germplasm

For QTL analysis, the researcher starts with inbred lines that differ for phenotypic expression of the trait of interest.

2.2. Molecular Marker Analysis

A detailed protocol for performing restriction fragment-length polymorphism (RFLP) analysis in maize (*Zea mays* L.) (easily adopted to other plants) is found at (http://www.maizemap.org/rflp_protocols.htm).

A detailed protocol for performing simple sequence repeat (SSR) analysis for maize is available from (http://www.maizemap.org/ssr_methods.htm).

For construction of genetic maps, one needs to obtain a copy of MAPMAKER/EXP, copyright 1992, The Whitehead Institute for Biomedical Research (*see* **Note 1**).

2.3. QTL Analysis

For single-factor analysis of variance, the most commonly used statistical software package is SAS, with information available at (http://www.sas.com).

For interval mapping (IM) and composite interval mapping (CIM), we use QTL CARTOGRAPHER available from (http://statgen.ncsu.edu/statgensoft_qtl.html). An excellent alternative software is PLABQTL ***(7)***, which is available

from the authors by anonymous file transfer protocol (ftp) as described in the paper cited.

3. Methods

3.1. Choice of Starting Germplasm

QTL are population-specific, therefore, the choice of starting germplasm will determine the QTLs identified. Researchers in crop improvement have the distinct advantage in that the starting materials can usually be inbred lines. The standard protocol is for the investigator to screen a large number of lines for the trait of interest (sometimes hundreds) to identify specific lines that differ widely for phenotypic values for the trait. While this approach practically guarantees QTL will be discovered and is fine for discovering the major QTL controlling traits, the investigator needs to consider whether the variability uncovered will be meaningful for subsequent goals of crop improvement. Many very negative alleles for traits may have already been eliminated from elite germplasm, and therefore, picking the poorest possible inbred for a trait may lead to discovery of QTLs that are fixed for the better allele(s) within breeders lines.

3.2. Population Structure

In searching for QTLs in crop plants, most populations involve crosses of inbred lines. This maximizes linkage disequilibrium and greatly simplifies identifying polymorphic markers and in assigning genotypes, because the original phase of all alleles can be defined directly from parental screenings. The appropriate population structure for a particular QTL experiment is influenced by a number of factors, including sample variance in measuring the trait, self-incompatibility with the species, and time available to develop the population before initiating the study.

3.2.1. Backcross Population

A backcross (BC) population is formed by making an initial cross followed by crossing the resulting F_1 plant with one of the original parents. This population type is often used in making populations between cultivated and wild species, where self-incompatibilities prohibit selfing plants. This population structure is easy to score and analyze, but because a full array of possible genotypes are not present and recombinant chromosomes are only captured from one side of the cross, BC populations are less powerful for QTL detection than F_2 or recombinant inbred line (RIL) populations. In addition, BC populations are particularly weak in detecting interactions between loci (epistasis). Another weakness of BC populations is the lack of replication in scoring phenotypes from individual plants.

3.2.2. F_2 Population

An F_2 population is developed by making the initial cross between two defined parental lines, usually inbred, followed by selfing of the F_1 plant to produce F_2 seed. One then derives both genotype and phenotype information from the F_2 plants. This is a powerful population structure for QTL detection, in that the full array of genotypes is present at each locus. This structure, therefore, permits estimates of additive and dominance effects at the QTL detected. The major weakness of an F_2 population is the lack of replication in scoring the phenotypes. The degree to which this is an issue depends on the trait under study (*see* **Note 2**). For traits such as yield or measures of insect resistance, sampling variance is too great to routinely use an F_2 population structure. This limitation can be overcome by selfing the F_2 plants to derive $F_{2:3}$ lines that can then be planted in replicated trails. Although genetic variability remains among the individuals within a specific $F_{2:3}$ line, this is generally of secondary importance relative to the reduction in phenotypic sampling error (*see* **Note 3**).

3.2.3. Recombinant Inbred Lines

Recombinant inbred lines are developed by repeated generations of selfing with each line starting from an individual F_2 plant. A standard level of selfing is six generations, at which the lines are homozygous for alleles at >98% of their loci. Because recombination events are captured between residual heterozygous regions during the cycles of selfing, an RIL population contains essentially the same number of crossovers as an F_2 population of similar size. A great benefit of an RIL population is that the genotypes of the lines are fixed, and therefore, once genotypes for a population are determined, the population can be used for any number of replications or measured for any number of different traits under specific growth conditions. Dominance values cannot be measured directly within an RIL population but, if desired, can be determined by developing BC populations with each of the original parent inbred lines (*see* **Note 4**). When one considers two unlinked loci in a RIL population, there will be four genotype classes of equal frequency. This distribution of genotype classes makes RIL populations very powerful for detecting two-locus epistatic interactions ***(8)***.

3.2.4. Population Size

There is always the same question at the start of any QTL project: "How large a population should I do?" This question is usually answered more by practical constraints than theoretical considerations of the power to detect QTLs. The main constraints are the cost of genotyping and the difficulty and cost of obtaining the phenotypic trait data. For a trait such as plant height in

maize, where gathering the phenotypic data is easy and inexpensive, constraint on the population size would come from the genotype cost. However, for a trait such as soybean cyst nematode resistance, which is labor- and time-intensive to score, the logistics of screening lines set a practical boundary on population size. Regardless, it is important for the researcher to consider how population size affects ability to detect and accurately estimate QTL effects. In a paper that is required reading for anyone contemplating a QTL experiment, Beavis ***(9)*** demonstrated that population sizes of 100–200 individuals, typical of most early QTL experiments, have very limited power to detect QTLs of small to moderate effects. He also demonstrated that for the QTLs detected, the QTLs' effects were generally greatly overestimated. With the advent of SSRs and other efficient marker systems, population sizes of at least 300 individuals should be the norm.

3.2.5. Phenotypic Data

The importance of accurate and relevant phenotypic data to the outcome of any QTL study cannot be overemphasized (*see* **Note 5**). The goal of QTL studies is to gain a genetic description of a particular trait or process. The removal of all possible environmental, sampling, and correlated trait variance is necessary to obtain the meaningful results.

3.3. Marker Analysis

Once the population is developed and trait data are obtained, the next step is to determine genotypes for all the individuals in the population.

3.3.1. RFLP

RFLP analysis was the critical technology breakthrough that enabled routine QTL analysis in crop plants ***(10)***. The RFLP method was adapted to essentially all crop species. The loci detected by RFLPs can be scored in a co-dominant manner and are transferable between populations. While RFLP analysis has been largely replaced by polymerase chain reaction (PCR) methods in the major crop plants, it remains a useful tool in secondary crops with limited genomic tools and is still the method of choice in establishing and utilizing cross-species syntenic relationships for QTL analysis.

3.3.2. SSR

SSRs, also commonly known in the animal–medical sciences as microsatellites, consists of direct tandem repeat of 2–6 nucleotides in length ***(11)***. Polymorphism is detected based on PCR primers designed to flank the repeat unit and to amplify a defined fragment containing the SSR. The number of repeat units evolves rapidly, leading to SSR markers exhibiting high poly-

morphism rates. Because they are PCR-based, SSR markers hold many advantages over RFLPs. The amount of DNA needed for assays is greatly reduced, there is no radioactivity involved in the assay, and because assays can be conducted in microtiter-plate format, SSR analysis is easier to conduct on a large scale than RFLPs. Because polymorphisms are detected as length differences, SSR markers provide co-dominant genotype information and, once mapped, are fully transferable between populations. There are two predominant formats for assaying SSR polymorphism. The first format is to resolve polymorphic fragments on agarose gels followed by staining with ethidium bromide. The advantages are low cost, simple equipment requirements, and ease of protocol setup and operation. The second major assay system is to use fluorescent-labeled primers and resolve PCR products on either gel-based or capillary-based automated DNA sequencers ***(12,13)*** (*see* **Note 6**).

Our protocol for SSR mapping with agarose gels is as follows:

1. Array DNA of the individuals in the mapping population into a 96-well stock plate at 10 ng/μL.
2. Transfer 50 ng DNA of each into wells of a 96-well thin-wall microtiter plate along with: 1× PCR buffer, 2.5 m*M* $MgCl_2$, 0.4 m*M* each dATP, dCTP, dGTP, and dTTP, 50 ng each forward and reverse SSR primers, 0.3 U AmpliTaq Gold® (Applied Biosystems) or Platinum *Taq* (Invitrogen), sterile water to a total vol of 15 μL, and overlay with drop of mineral oil.
3. The PCR program is: 10 min at 95°C; then one cycle of 95°C for 1 min, 65°C for 1 min, 72°C for 90 s; 1°C decrement in annealing temperature per cycle until annealing temperature is 55°C (this takes 10 cycles); then 30 cycles of 95°C for 1 min, 55°C for 1 min, 72°C for 90 s.
4. Resolve SSR products on 3 to 4% super-fine resolution (SFR) agarose gels (Amresco) by electrophoresis at 115 V or 2 to 3 h depending on size of SSR products. The SFR gels are made in 1× TBE (90 m*M* Tris-borate, 1 m*M* EDTA, pH 8.3). Ethidium bromide (1 mg/mL) is added immediately before gels are poured. Used gels can be melted by microwave and reused many times (>40 times).
5. Photograph gel with charge-coupled device (CCD), we find it easiest to post an electronic copy of the gel image to an NT network to be scored at a later date.

3.3.3. Multilocus PCR

Although both amplified fragment-length polymorphism (AFLP) ***(14)*** and random-amplified polymorphic DNA (RAPD) ***(15)*** methods generate genotypes at multiple loci per PCR, their utility for QTL analysis has been limited by lack of transferability between populations, making it difficult to complete full genome coverage. Also, AFLP and RAPD markers are scored as dominant markers and, therefore, are of lower information content if mapping in population structures with heterozygous individuals.

3.3.4. Single Nucleotide Polymorphism

The most common type of polymorphism in any species is the single nucleotide polymorphism (SNP). SNPs may occur in crop plants as frequently as 1 out of 100 nucleotides in maize ***(16)***. There is currently intense activity to define SNPs in the major crop species and to develop high-throughput assays. Because of the potential for automation, the development of SNP assays and providing SNP genotyping services is an area of intense commercial activity. It is anticipated that SNPs will soon become the most common genotyping method in the major crops, and because of the economy of scale, it may soon be cost efficient to "contract out" genotyping for a QTL project rather than conduct it within one's laboratory. Currently, there are a large number of methods for SNP genotyping. In our laboratory, we perform a multiplex primer extension assay using the ABI SnapShot™ Kit (Applied Biosystems) as per manufacturer's recommendations (www.appliedbiosystems.com). As with fluorescent SSRs, the products of the SnapShot kit can be resolved on either gel-based or capillary-based DNA sequence systems.

3.3.5. Marker Summary

For the investigator entering QTL analysis today, the most practical marker system will usually be SSRs. Information on the public SSR resources for major crops is readily available through the species-specific genetic databases (*see* **Note 7**). In the near future, SNP genotyping should become a cost-effective alternative.

3.4. QTL Analysis

3.4.1. Map Construction

As will be discussed in the next section, both IM and CIM require a genetic map of the experimental population for analysis. Clearly, the quality of one's results are dependent on the quality and completeness of one's genetic map. Even if the investigator is going to conduct single-factor analysis, which does not require a map, constructing linkage maps for the regions of interest allows for the examination of chromosome coverage and serves as a check on genotype quality (*see* **Note 8**). Although the program is old and no longer actively supported, the most commonly used program for genetic map construction remains MAPMAKER/EXP. This program supports maps for BC, F_2, and RIL populations (*see* **Note 9**).

The steps for constructing a genetic map with MAPMAKER/EXP are as follows:

1. Establish a .raw import file with all your genotype and trait data as outlined in the

manual. Start program using the *prepare* command. Set *mapping function* command to Haldane's, use *print names on* command to track results by marker names instead of number, use *photo on* command to make a text copy of all subsequent steps.
2. Execute the *group* command to use two-point analysis to form linkage groups. We start with the default setting of likelihood of odds (LOD) 3.0 and 50 cM for inclusion into a group. If more groups than chromosomes are obtained, one can extend (with caution) to 60 or 65 cM to see if multiple initial groups will merge.
3. Look at the markers present in each linkage group to determine which groups correspond to which chromosomes–linkage groups. Use *make chromosome* command to define chromosomes and *anchor* command to designate groups to chromosomes.
4. Use the *order* command to resolve marker order within a chromosome. Examine the order to see if chromosomes are "right side up." If not, use *sequence* command to reverse order of markers. Once you have markers in proper order, use the *framework* command to make a "frameworked" chromosome. You must have a framework chromosome for each named chromosome for MAPMAKER/QTL or QTL CARTOGRAPHER to conduct a full genome search for QTL.
5. Use *error detection on* and *genotypes* commands to get a visual output for all the cross-overs within your population. Examine the output for markers with high numbers of double recombinations around specific markers that should then be checked for genotyping errors.
6. Resolve genotyping errors discovered by **step 5** and redo **steps 1–4** to redraw an accurate genetic map.

The question always comes up as to how dense a map is desired for QTL experiments. In a standard QTL experiment of 200–300 individuals, a reasonable marker density for initial analysis is about 15 cM between markers. Little additional power is gained in increasing marker density beyond this level, because there are then not enough recombination events to further refine QTL position with additional markers ***(17)***.

3.4.2. QTL Analysis

There are three basic methods of QTL analysis: single marker, IM, and CIM. In single-marker analysis, tests are performed separately for each marker to test if the genotype class is significant for the phenotype values for the trait of interest. This method cannot provide an accurate estimate of QTL position, and therefore, estimates of QTL effects are confounded by the actual position of the QTL relative to the marker. The introduction of the MAPMAKER/QTL software ***(18)*** partially solved the issue of QTL position by using the linkage map for the QTL experiment and maximum likelihood estimates to scan a chromosome for the position of highest probability for the QTL. With the position

known, better estimates of QTL effects are possible. However, the positions and genetic effects of the QTLs detected by IM can be confounded by the presence of other linked QTLs or by nonrandom segregation of other QTLs in the population. These two limitations with IM are addressed with CIM ***(7,19)***. In CIM, significant markers in the population, identified by regression analysis, are used as cofactors in determining maximum likelihood estimates for QTL, and therefore, effects of other QTL are taken into account. This often allows separation of a single broad region of QTL identified by IM into two QTLs by CIM. The two most commonly used implementations of CIM in public software are QTL CARTOGRAPHER and PLABQTL. Our laboratory makes extensive use of the QTL CARTOGRAPHER, which has a clear and detailed on-line manual (http://statgen.ncsu.edu/qtlcart/cartographer.html), that is often updated.

The basic steps in performing CIM using QTL CARTOGRAPHER are as follows:

1. Build a linkage map with MAPMAKER/EXP as detailed in **Subheading 3.4.1**. The .raw and .maps files from MAPMAKER/EXP serve as the starting files for QTL CARTOGRAPHER.
2. QTL CARTOGRAPHER is operated as a set of subprograms, each with a set of lines that can be set by the user. Unless specified, we start with the default setting for lines. We use the *Rcross*, *Rmap*, *Srmapqtl*, *Zmapqtl*, and *Zmapqtl* with permutations programs to conduct a standard CIM analysis.
3. Import the genotype, trait, and map data (.raw and .maps) into QTL CARTOGRAPHER with the routines *Rcross* and *Rmap*. Set a name for the project for all the output files using the "change Filename stem" line on the first program you execute. All subsequent output files will be named with that project name for easy tracking. For the UNIX® system, it often requires executing the *Rcross* and *Rmap* programs twice before the data are accepted.
4. Select cofactors for use in CIM forward–backward regression with the program *SRmapqtl* (*see* **Note 10**). *SRmapqtl* will need to be executed for each trait under study. This can be done in one step by setting the trait number at one greater than the total number of traits present in the .raw file. The results of the *SRmapqtl* analysis are presented in the .sr file.
5. Conduct CIM with *Zmapqtl* under Model 6 (set "model" line to 6) using all cofactors identified by *SRmapqtl*. This is done by setting the "number of cofactors" line to a number greater than the number of cofactors that were detected by *SRmapqtl*. Set chromosomes to analyze to "0" to conduct a full genome scan. The output from *Zmapqtl* is in an output file named .z
6. Execute *Eqtl* to summarize QTL results for ease to finding significant regions.
7. Conduct CIM with *Zmapqtl* as in **step 5** with permutations set at 1000 to determine genome-wise thresholds as described below.

Once one has conducted CIM, the investigator is faced with the decision of choosing the threshold for declaring the presence of a QTL. In the early days of QTL analysis, the thresholds were generally arbitrarily set. This often resulted in setting thresholds dependent on the outcome of the analysis, rather that the other way around, leading to a large number of reported QTLs of questionable validity. Sanity was introduced to this problem with the implementation of permutation or bootstrap protocols for empirically determined threshold levels specific to the population under study ***(20,21)***. We prefer permutation analysis as performed in **step 7** above, because it maintains all genetic map and trait population parameters. In permutation analysis, the trait values are randomly reassigned to individuals, and CIM is performed, each reiteration giving a value for the highest false QTL detected. We have conducted this analysis for a number of populations and traits for maize, and the $P = 0.05$ experiment-wise threshold is most often in the range of log LOD ratio of 3.3–3.6, much higher than the thresholds used for most QTL studies reported in the literature (*see* **Note 11**).

3.5. Cloning QTL

We have entered the era in which genes, identified first as QTLs, are being cloned and characterized. I will discuss three examples that show the power of using QTL analysis for gene discovery. These examples are cloning of *fw2.2* in tomato ***(2)***, and *Hd1* ***(3)***, and *Hd6* ***(22)*** in rice. In all three examples, the steps from QTL to gene are similar.

1. "Mendelize" the QTL by backcrossing to develop a near isogenic line (NIL), where only the region of the QTL is heterozygous, removing the genetic variation caused by other QTL.
2. Develop a fine-structure map in a large (1000–3000) population of the NIL.
3. Transfer the genetic fine-structure map covering the QTL onto a physical map.
4. Sequence to identify the candidate gene(s) within the defined region.
5. Confirm the candidate gene as the basis of the QTL by transformation of the gene into a genetic background that allows complementation of phenotype.

The *fw2.2* QTL was identified as a major QTL for fruit weight in tomato. The *fw2.2* gene was cloned by the process outlined above and was demonstrated to encode a RAX family protein, which is a regulatory factor. The *Hd1* locus in rice was identified as the major QTL controlling heading date in a subspecies cross in rice. *Hd1* was cloned in a manner similar to *fw2.2*. The *Hd1* gene encodes a zinc-finger domain protein with similarity to the photoperiod sensitive gene CONSTANS in *Arabidopsis*. With its zinc-finger domain, *Hd1* is presumably a transcription factor. The isolation of the gene for the heading date QTL *Hd6* demonstrates that QTLs with small effects may also be targets

for characterization. *Hd6* was not detected in the initial F_2 population that identified *Hd1*, but was detected in chromosomal substitution line analysis of the same cross. In chromosomal substitution line analysis the genome of one genotype is placed one segment at a time into another genotype to form a series of NILs. Despite the small effect of *Hd6* on heading date, the researchers were able to develop a detailed fine-structure map in 2807 individuals and isolate the gene as for *Hd1*. *Hd6* encodes a subunit of protein kinase CK2 and is believed to function in the signal transduction pathway leading to flowering.

3.6. Summary: The Role of QTL Analysis in Gene Discovery

In the three examples cited above, the bases of the QTLs were regulatory or signal transduction proteins. Frary et al. ***(2)*** reported that the mRNA level of *fw2.2* was too low to detect by Northern analysis in any tissue and could only be characterized by reverse transcription-PCR (RT-PCR). Yano et al. ***(3)*** reported that at the time of cloning, there was no cDNA for *Hd1* in the EST collections of rice. By using candidate gene approaches, transcription factors have also been shown to be the basis of QTL for maize domestication ***(23)*** and corn earworm resistance in maize silks ***(24)***. By extension, it is reasonable to assume that transcription and signal transduction factors underlay many of the QTLs controlling plant growth, development, productivity, and response to biotic and abiotic stresses. Because transcription factors have low mRNA levels, or may be expressed in only very specific cells or stages of development, they are generally underrepresented in EST collections and are, therefore, missing from currently available microarrays for many crop species. To be detected as significant in microarray analysis, the expression level of a gene must vary at least two- or three-fold or greater. Though changes in the level of expression of genes for enzymes will often vary by this amount, the changes in levels of regulatory genes may be subtle, particularly if the changes in mRNA levels occur in a limited number of cells. The basic difference between genomic approaches, such as microarray analysis and QTL analysis, is that, while microarrays identify genes that respond, QTL analysis identifies genes that regulate. Even with a full genome sequence, such as in *Arabidopsis*, if all the transcription factors were identified, QTL analysis is a useful approach to linking function with specific genes. It is this power of QTL analysis to identify the key regulatory components that requires the integration of QTL analysis with genomics.

Jansen and Nap ***(25)*** have proposed a framework for the integration of QTL analysis and genomics that they have termed "genetical genomics." Their strategy calls for conducting microarray expression analysis or proteomic analysis in the context of structured segregating populations, if the changes in mRNA or protein levels can be mapped as traits. The biochemical pathways involved

in trait expression can be identified by the changes in RNA or proteins for the enzymes involved, and the regulatory factors controlling those traits are mapped as the QTL for mRNA or protein levels. In this manner, clues to both the biochemical basis and regulatory factors controlling complex agronomic traits can be discovered.

4. Notes

1. The MAPMAKER software is no longer actively supported by any institution or individual, however, it can be freely redistributed, and copies can be readily obtained from colleagues.
2. In our laboratory, we have made extensive use of F_2 populations in QTL studies of maysin concentration in maize silks. Because we can accurately measure maysin levels on a single-plant basis, the F_2 population structure is very appropriate ***(24)***.
3. One adjustment made in estimating gene effects in an $F_{2:3}$ population is that only half of the dominance deviations are detected, therefore, dominance estimates derived for an $F_{2:3}$ population need to be doubled to reflect true allele relationships.
4. I think this approach has been underutilized to date, particularly in crops like maize where hybrids are grown, and understanding dominance effects and measuring trait responses in the presence of heterosis are agronomically relevant conditions.
5. While this statement may seem obvious, I feel that many published QTL studies suffer from poor trait data or from measuring traits in a way that correlated traits confound one's ability to measure primary vs secondary effects.
6. Depending on the manufacturer's equipment, up to four different dyes can be multiplexed. Multiplexing is also accomplished by product size. Multiplexing multiple dyes with multiple sizes per dye can allow from 9–12 SSRs to be resolved per assay. Because of PCR competition, most investigators perform separate PCRs and pool samples only for analysis on the automated sequencer. While combinations of primer that can be multiplexed in the PCR can be identified, this normally requires extensive troubleshooting before actual genotyping can begin.
7. A list of the Web sites for the plant genetic databases is available from either (www.ukcrop.net) or (www.agron.missouri.edu/bioservers.html).
8. Gene order and approximate map distances are expected to be conserved within a species, and deviations of order or very different map distances from standard well-defined species maps may highlight markers with genotype problems for careful examination by the researcher.
9. One of the strongest features of the program is the ability to set statistical thresholds for declaring marker order. For QTL populations of 200–300 individuals, high standards for inclusion into a chromosome and map order should be used. The *error detection* and *genotype* functions are very useful for identifying problem markers. The data for the genotypes for these markers can then be reexamined or redetermined to add confidence to map quality.

10. We have found for our population and trait structures that setting the thresholds to include and retain markers to 0.01 [(p(Fin) = p(Fout) = 0.01] usually results in a reasonable number (3–8) of significant regions to use as cofactors.
11. This is a very strict threshold, and some QTL with small effects are missed. For purposes of identifying all QTLs, a lower threshold may be initially used, however, QTLs identified should be confirmed by progeny analysis or through the development of NILs.

References

1. Paterson, A. H. (ed.) (1998) *Molecular Dissection of Complex Traits.* CRC Press, Boca Raton.
2. Frary, A., Nesbitt, T. C., Frary, A., et al. (2000) *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289,** 85–88.
3. Yano, M., Katayose, Y., Ashikari, M., et al. (2000) *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene CONSTANS. *Plant Cell* **12,** 2473–2483.
4. Lander, E. S. and Schork, N. J. (1994) Genetic dissection of complex traits. *Science* **265,** 2037–2048.
5. Churchill, G. A. and Doerge, R. W. (1998) Mapping quantitative trait loci in experimental populations, in *Molecular Dissection of Complex Traits* (Paterson, A. H., ed.), CRC Press, Boca Raton, pp. 31–42.
6. Kearsey, M. J. and Farquhar, A. G. L. (1998) QTL analysis in plants; where are we now? *Heredity* **80,** 137–142.
7. Utz, H. and Melchinger, A. E. (1996) PLABQTL: a program for composite interval mapping of QTL. *J. Agric. Genomics* **2,** 1–4.
8. Orf, J. H., Chase, K., Jarvik, T., et al. (1999) Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. *Crop Sci.* **39,** 1642–1651.
9. Beavis, W. D. (1998) QTL analyses: power, precision, and accuracy, in *Molecular Dissection of Complex Traits* (Paterson, A. H., ed.), CRC Press, Boca Raton, pp. 145–162.
10. Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–199.
11. Tautz, D., Trick, M., and Dover, G. A. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322,** 652–656.
12. Mitchell, S. E., Kresovich, S., Jester, C. A., Hernandez, C. J., and Szewc-McFadden, A. K. (1997) Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. *Crop Sci.* **37,** 617–624.
13. Cregan, P. B., Jarvik, T., Bush, A. L., et al. (1999) An integrated genetic linkage map of the soybean genome. *Crop Sci.* **39,** 1464–1490.
14. Vos. P., Hogers, R., Beeeker, M., et al. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23,** 4407–4414.
15. Williams, J. G. K., Reiter, R. S., Young, R. M., and Scolnik, P. A. (1993) Genetic

mapping of mutations using phenotypic pools and mapped RAPD markers. *Nucleic Acids Res.* **21,** 2697–2702.
16. Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5,** 94–100.
17. Darvasi, A., Weinreb, A., Minke, V., Weller, J. I., and Soller, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134,** 943–951.
18. Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988) Resolution of quantitative trait into Mendelian factors by using a complete map of restriction fragment length polymorphisms. *Nature* **335,** 721–726.
19. Zeng, Z.-B. (1994) Precision mapping of quantitative trait loci. *Genetics* **136,** 1457–1468.
20. Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.
21. Doerge, R. W. and Churchill, G. A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142,** 285–294.
22. Takahashi, Y., Shomura, A., Sasaki, T., and Yano, M. (2001) *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. *Proc. Natl. Acad. Sci. USA* **98,** 7922–7927.
23. Doebley, J. and Lukens, L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell* **10,** 1075–1082.
24. McMullen, M. D., Byrne, P. F., Snook, M. E., et al. (1998) Quantitative trait loci and metabolic pathways. *Proc. Natl. Acad. Sci. USA* **95,** 1996–2000.
25. Jansen, R. C. and Nap, J.-P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.* **17,** 388–391.

10

Transposon Tagging Using *Activator (Ac)* in Maize

Thomas P. Brutnell and Liza J. Conrad

Summary

The transposable element *Activator (Ac)* has been used in several plant species as a tool for gene isolation and characterization. However, it has not been widely utilized in its native host maize, in part, because of a relatively low germinal transposition rate. The propensity of *Ac* to move to linked sites provides an opportunity to overcome this limitation when *Ac* elements are distributed at regular intervals throughout the genome. This chapter details the use of such a system in maize through simple genetic manipulations. A detailed protocol is also provided to clone DNA flanking *Ac* insertions.

Key Words

Ac, maize, transposon, IPCR, tagging, mutagenesis

1. Introduction

Activator (*Ac*) and *Dissociation* (*Ds*) were the first transposable elements discovered ***(1)*** and have been intensively studied using both classical genetic and molecular genetic techniques (for an excellent review of *Ac/Ds* biology *see* ref. ***2***). Despite 50 yr of research and widespread use in transgenic systems ***(3–7)***, the *Ac/Ds* family has not been used extensively in maize as a tool for gene isolation, particularly in recent years. This can largely be attributed to two important characteristics of *Ac/Ds* regulation. First, the autonomous element *Ac* transposes at a rate nearly 50- to 100-fold lower than the widely utilized *Mutator* family of transposable elements ***(8)***. Consequently, only 2–4% of the progeny inherit a newly transposed *Ac* or *Ds* element (***[9]*** and Brutnell and Conrad, unpublished results). This contrasts with the *Mutator* system, in which up to 10 new transposition events may be inherited per progeny (Bruce May, Cold Spring Harbor Laboratory, personal communication). A second

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

characteristic that has limited the utility of *Ac/Ds* in transposon mutagenesis programs is the tendency of these elements to insert at genetically linked sites. In studies of the *p* and *bz* loci, 60% of *Ac* transpositions were to genetically linked sites ***(10–12)***. Of these, the majority were to sites within 10 cM of the donor element. Thus, the preferential short-range transposition of *Ac*, together with the low forward transposition rate of *Ac* or *Ds*, has greatly limited the use of this transposable element family in nontargeted random mutagenesis experiments.

Despite these limitations, there are a number of features of *Ac/Ds*, which make them particularly attractive for use in gene tagging and characterization experiments ***(13)***. For one, the relatively low copy number of active *Ac/Ds* elements in the maize genome offers some advantages. Selection for *Ac/Ds* excision or *Ac* transposition will often result in the segregation of a single new element in the following generation. In contrast, the high copy number of *Mu* elements can make segregation analysis difficult ***(14)***. Furthermore, the high mutational load associated with active *Mu* lines, makes it highly likely that a *Mu* insertion in the gene of interest will be present in a genome containing multiple *Mu* insertions. Thus, when characterizing *Mu*-induced alleles, it is advisable to introgress the insertion alleles into a standard inbred line prior to extensive phenotypic characterization. Because *Ac* is maintained at low copy number in the genome, genetic and molecular characterizations are greatly simplified.

Although the propensity for linked transpositions can limit the use of *Ac/Ds* in random mutagenesis programs, it can be exploited to generate multiple alleles of a closely linked gene. Perhaps the most elegant use of *Ac* in regional mutagenesis has been in studies of the *p* locus where several hundred *Ac*-induced alleles have been generated through localized transposition events ***(15–18)***. These studies highlight the utility of *Ac* not only in gene tagging, but also in fine-scale genetic mapping. *Ac* insertional mutagenesis at the *p* locus has revealed promoter sequences, intron–exon boundaries and enhancer sequences >4.0 kb upstream of the start of transcription. In addition to providing detailed structural information for the gene of interest, *Ac*-/*Ds*-induced alleles can also be used for site-directed mutagenesis. Because *Ac/Ds* excision is often imprecise, stable frameshifts, nonsense and missense mutations can be generated from unstable alleles. These derivative alleles can be used to verify the identity of a tagged gene (e.g., ***19***) or to generate an allelic series for detailed phenotypic characterizations.

Although the *Ac* family potentially affords many advantages in gene tagging, no systematic study has been performed in maize, examining the frequencies of *Ac* insertion into genetically linked target loci. Nevertheless, two general strategies have been described for utilizing *Ac*/*Ds* as insertional

mutagens in maize ***(20)***. In one strategy, transposition events are recovered as single kernel events following excision of *Ac* or *Ds* from a reporter gene. As demonstrated at the *p* locus, *Ac* can be recovered in two-thirds of the gametes when an excision assay is used to screen for transposition events. However, when single kernel events are selected, there is only a 50% chance that the gamete carrying the new *tr-Ac/Ds* will be recovered. Thus, only 33% of the kernels will carry a new *tr-Ac* or *tr-Ds*. Nevertheless, this strategy has been successfully utilized to clone the *Indeterminate1* (*Id1*) gene of maize ***(21)***. To tag *Id1*, Colosanti and colleagues utilized a *Ds* element resident at the *bz2* locus, 1 cM from *Id1*. Approximately 600 fully colored kernels (germinal *Ds* revertants) were selected, and one *Ds*-induced allele was recovered ***(21)***. Assuming the frequency of *Ds* elements recovered is similar to the number of *Ac* elements recovered following germinal excision events, only 200 of the 600 kernels selected (33%) would have been expected to carry a *tr-Ds* element. Unfortunately, as only one allele was recovered, the frequency of *Ds* insertion cannot be ascertained. An *Ac* excision assay was also used in an insertional mutagenesis of the *R-nj* gene of maize. In this screen, an active *Ac* was positioned within approx 20 cM of the *r* locus through the use of a reciprocal translocation ***(22)***. Approximately 78,000 kernels were screened, and four mutable *R-nj* alleles were recovered. Again, only one *Ac*-tagged allele was characterized molecularly, thus, it is not possible to accurately gauge the frequency of *Ac* insertion.

In another strategy for *Ac* tagging, *Ac* transposition events can be selected on the basis of increased *Ac* copy number ***(20)***. Increases in *Ac* copy number results in delayed transposition of *Ac* and *Ds* and, consequently, smaller revertant sector sizes ***(23)***. This feature of *Ac*, known as the negative dosage effect, provides a convenient assay to monitor the copy number of *Ac* in the genome. One of the advantages of this strategy is that *Ac*s, located anywhere in the genome, can be exploited, as selection utilizes a *Ds* reporter and not an *Ac* excision assay. Another advantage is that with a suitable *Ds* reporter, both endosperm and embryo can be monitored to ensure concordance of embryo and endosperm genotypes. This is particularly important when the increase in *Ac* copy number results in a colorless rather than a finely spotted aleurone. In these instances, the ability to monitor *Ac* copy number in the scutellum (derived from the embryo) provides a convenient assay to ensure that a *tr-Ac* was inherited in embryonic tissues (*see* **Fig. 1**). General strategies for gene tagging with *Ac* and precautions that must be taken when maintaining *Ac* stocks have been described elsewhere ***(20,24–26)***.

As the frequency of *Ac* transposition is relatively low, we have devised two strategies for gene tagging that exploit the negative dosage effect of *Ac* to select for transposition events. If existing mutant alleles of a gene of interest are avail-

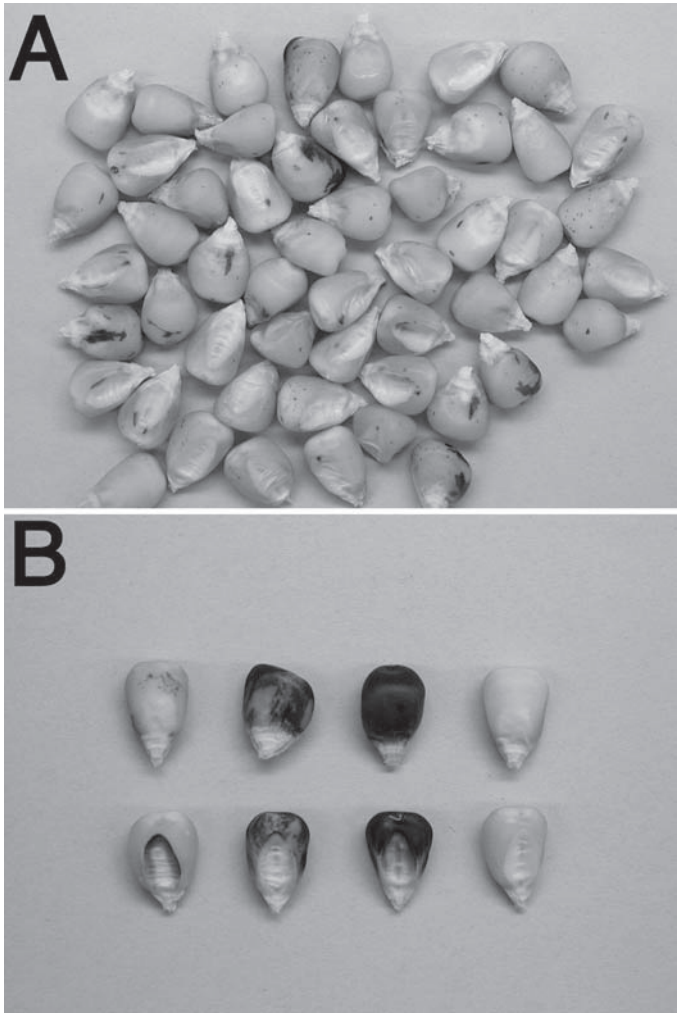


Fig. 1. Typical kernel phenotypes generated from self-pollinated *tr-Ac/tr-Ac*, *r-sc:m3/r-sc:m*3 lines. **(A)** Expected variegation pattern observed with lines homozygous for an *Ac* insertion and the *r-sc:m3* reporter. Notice that a finely spotted variegation pattern can be seen in the aleurone and scutellum tissues in many of the kernels. **(B)** Off-types shown representing (from left to right) germinal excision of *Ds* reporter, resulting in a fully colored scutellum, coarse spotting pattern in aleurone, indicative of a heterozygous (*Ac*/+) individual, fully colored kernel, indicative of a premeiotic *Ds* excision event, and a completely colorless kernel, indicative of a loss of *Ac* activity.

able and condition a nonlethal mutant phenotype, then a directed mutagenesis may provide the best chance of recovering *Ac*-induced alleles. In a directed mutagenesis, insertion alleles can be screened in the F1 generation, thus saving the time and labor necessary to create an F2 population. As the self-pollination of F1 plants will likely be the rate-limiting step of a random mutagenesis, a directed tagging experiment will likely mean that a larger F1 population can be generated increasing the chances of success. One obvious disadvantage of a directed tagging experiment, is that potential *Ac*-induced alleles will be recovered as single kernel events. Thus, care must be taken to ensure that any mutants identified can be propagated through either self-pollination or out-crossing. An additional limitation is that screens will most likely be performed in a hybrid background. Thus, modifier loci, introduced from the W22 parent, could potentially mask the phenotypic effects of mutant alleles maintained in a different genetic background. Finally, selections for new transposition events will need to be performed using the *Ds* reporter at the *r* locus. An *r* allele that conditions a colored aleurone introduced from line carrying the reference allele will prevent the use of the *r-sc:m3* reporter. However, as all of the *Ac* lines currently under development are maintained as *r-sc:m3* homozygotes, any recessive *r* allele introduced from the reference line will permit the selection of transposition events.

If existing mutant alleles condition a lethal or sterile phenotype, or existing alleles are maintained in a genetic background that is incompatible with the *Ds* reporter, then a nondirected regional mutagenesis (random mutagenesis) will most likely be the optimal strategy to utilize. In a regional mutagenesis, new transpositions are selected in the F1 generation following a testcross of plants homozygous for an *Ac* insertion by plants homozygous for the *Ds* reporter. These F1 plants are then self-pollinated to generate a segregating F2 population for kernel, seedling, or mature plant screens. One clear advantage of the regional mutagenesis is that all lines are maintained in a uniform genetic background, permitting detailed phenotypic comparisons of any recovered alleles. Of course, one can envision variations on either of these general strategies that may improve the chances of success depending on the mutant phenotype. Experiments are currently in progress in our laboratory to examine the advantages and disadvantages of both methods of *Ac* mutagenesis.

2. Materials

2.1. Transposon Tagging

Our laboratory is currently distributing *Ac* throughout the genome for use in targeted mutagenesis (http://bti.cornell.edu/Brutnell_lab2/Projects/Tagging/BMGG_pro_tagging.html). The goal of our program is to create approx 200

lines each containing an *Ac* element at a unique position within the maize genome and ideally separated by 20 cM intervals. Thus, most genes in the maize genome should be within 10 cM of an active *Ac*. All lines generated are maintained in a color-converted W22 population ***(27)*** and utilize the *Ds* reporter *r-sc:m3* ***(28)*** to monitor *Ac* activity. Each line is homozygous for an *Ac* insertion that has been positioned on one of two public recombinant inbred populations, the IBM94 population developed at the University of Missouri (http://www.maizemap.org/resources.htm) or the BNL96 population developed at the Brookhaven National Laboratory (http://burr.bio.bnl.gov/acemaz.html) (*see* **Note 1**). Detailed mapping data can be found on our laboratory Web site listed above.

1. Determine which *Ac* line(s) are within 10–20 cM of the gene-of-interest and request that seed stock. All seed stocks are deposited at the Maize Genetics COOP in Urbana, IL and are freely distributed. Genetic resources are limited, so only 10 kernels are distributed/line/laboratory (*see* **Note 2**).
2. Seed stocks must be propagated by self- or sib-pollination to generate a sufficient seed stock for large-scale mutagenesis.
3. The *Ac* donor lines do not condition an obvious mutant phenotype, so care must be taken to ensure that the *Ac* is resident at its mapped location. It is recommended that two selection criteria be used to ensure the genetic uniformity and integrity of the lines following self-pollination of the seed stocks:
   a. Kernels should display relatively few colored aleurone sectors, indicative of a homozygous *Ac* insertion, and off-type kernels should be discarded (*see* **Fig. 1, Notes 3** and **4**).
   b. Progeny kernels from each self-pollinated ear should be genotyped prior to mutagenesis (*see* **Note 5**). Genotyping should be performed through DNA blot analysis and a polymerase chain reaction (PCR) assay as shown in **Fig. 2**.

2.2. DNA Blot Analysis

1. 3–5 μg Genomic DNA.
2. Appropriate restriction enzyme(s) (*Eco*RI or *Pst*I).
3. 0.8% Agarose in Tris-acetate EDTA (TAE) (standard electrophoresis-grade Low EEO agarose [Fisher Scientific, cat. no. BP160–100]).
4. DNA ladder VII, digoxygenin (DIG)-labeled (Roche Molecular Biochemicals).
5. Hybond® N^+ nylon membrane (Amersham Pharmacia Biotech).
6. DIG Probe Synthesis kit (Roche Molecular Biochemicals).
7. DIG Easy Hyb Buffer (Roche Molecular Biochemicals).
8. Buffers required for DNA Blot detection:
   a. Low stringency wash: 2× standard saline citrate (SSC)/0.1% sodium dodecyl sulfate (SDS).
   b. High stringency wash: 0.5× SSC/0.1% SDS.
   c. Maleic acid buffer: 0.1 *M* maleic acid, 0.15 *M* NaCl, pH 7.5.

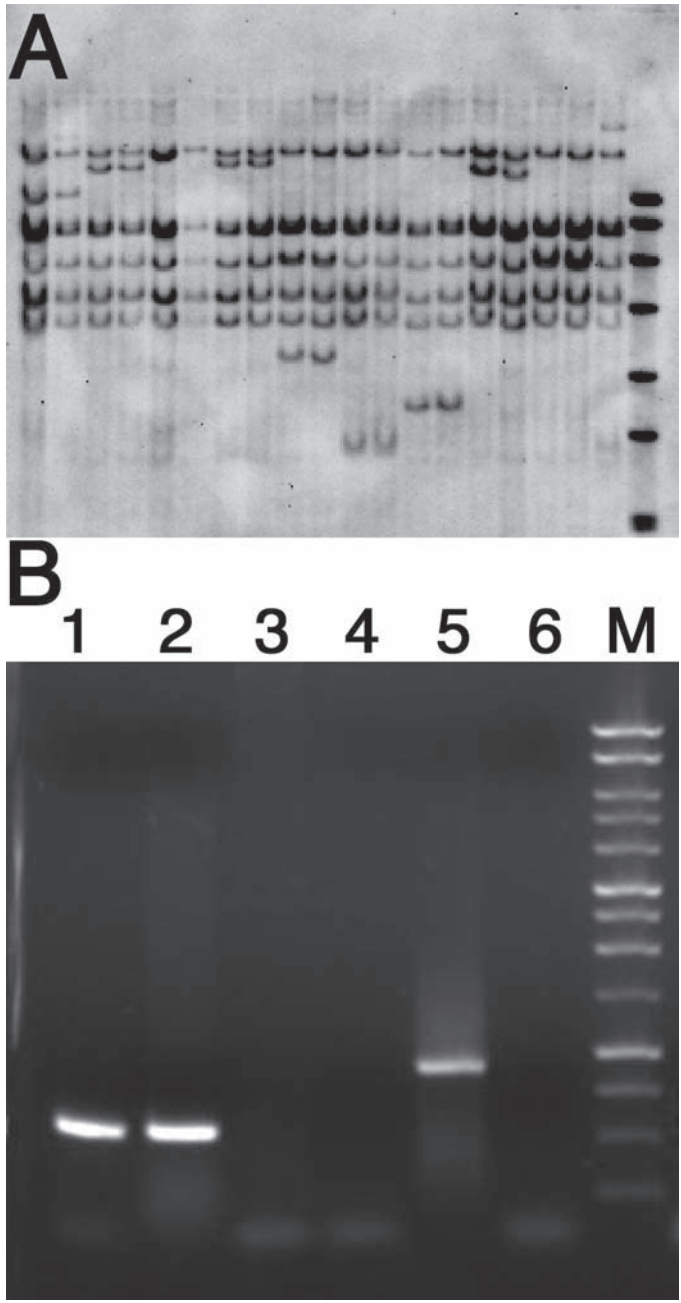


Fig. 2. Genotyping *Ac* lines. **(A)** DNA blot analysis was performed on several pairs of lines carrying independent *Ac* insertions. Genomic DNA was digested with *Eco*RI, fractionated on a 0.8% agarose gel, and transferred to nylon membranes. Blots were probed with a 900-bp *Eco*RI-*Hin*dIII internal fragment of *Ac*. The nontransposable *Ac*-homologous or cryptic fragments can be seen as common bands present in each

d. Blocking buffer: 10% casein, 0.1 *M* maleic acid, 0.15 *M* NaCl.
e. Washing buffer: 3% Tween® 20, 0.1 *M* maleic acid, 0.15 *M* NaCl.
f. Detection buffer: 0.1 *M* Tris-HCL, 0.1 *M* NaCl, pH 9.5.
g. Stripping buffer: 0.2 *M* NaOH, 1% SDS.

9. X-ray film or chemilumenescent detection system.

2.3. Inverse PCR (IPCR)

1. 15–20 μg Genomic DNA containing *Ac* insertion.
2. Appropriate restriction enzyme (*Eco*RI or *Pst*I).
3. 0.8% agarose in TAE (standard electrophoresis grade Low EEO agarose).
4. GeneClean® III kit (BIO 101).
5. T4 DNA ligase 3 U/μL 10× ligation buffer (Promega).
6. *Taq* DNA polymerase (Promega) (in storage buffer B, 10× buffer with 15 mM $MgCl_2$).
7. dNTP's (Promega) diluted to 2 mM in distilled water (dH_2O).
8. *Ac*-specific primers (listed in **Table 1**) at 10 μM concentration in dH_2O.
9. Dimethyl sulfoxide (DMSO) (Sigma)
10. Gel Extraction kit (Qiagen).
11. pGem®T-Easy Vector System I (Promega).
12. Nucleotide Removal kit (Qiagen).

3. Methods

3.1. Strategies for Gene Tagging

3.1.1. Random Mutagenesis

1. Transpositions are generated from a donor *Ac* (*d-Ac*) line that is positioned near (within 10–20 cM) a target locus following a test-cross of the homozygous *Ac* line by the *Ds* reporter line (*see* **Fig. 3** and **Note 6**).
2. New transpositions are selected as finely spotted F1 kernels among coarsely spotted siblings (*see* **Notes 7** and **8**).

Fig. 2. *Continued*
sample. The active *Ac* can be readily discerned in most lanes as a unique fragment present in the two individuals that were genotyped for each line. **(B)** PCR analysis to molecularly confirm *Ac* insertion site. PCRs were performed with *Ac*- and flanking sequence-specific primer pairs (primer sequences available on laboratory Web site). Lanes 1 and 2 show two individuals homozygous for an *Ac* insertion (*tr-Ac:mon0004*). Lanes 3 and 4 represent the donor *Ac* and *Ds* tester lines, from which the transposition event was originally generated. Lane 5 is a positive control using a plasmid control, and lane 6 is the negative control. Lane M is the 1 kb ladder. A predicted 530-bp amplification product is observed in both individuals carrying the *tr-Ac:mon0004* allele, representing an *Ac*-flanking sequence junction fragment. As this *Ac* insertion is not present in either parental line, no amplification products are observed.

Table 1
PCR Primer Pairs

Restriction enzyme	1st round PCR	2nd round PCR
*Eco*RI (2.0 kb)	JGp2:CCGGTTCCCGTCCGATTTCG	TBp32:CAAACATACCTGCGAGGATCAC
	TBp43:GAATTTATAATGATGACATGTACAAC	JGp3:ACCCGACCGGATCGTATCGG
*Eco*RI (2.5 kb)	TBp35:GTCGGGAAACTAGCTCTACCG	TBp34:ACCTCGGGTTCGAAATCGATCGG
	TBp42:GGCTGTAATTGCAGGAACAATTG	TBp37:TAATGAAGTGTGCTAGTGAATGTG
*Pst*I	TBp35:GTCGGGAAACTAGCTCTACCG	TBp34:ACCTCGGGTTCGAAATCGATCGG
	JGp2:CCGGTTCCCGTCCGATTTCG	JGp3:ACCCGACCGGATCGTATCGG

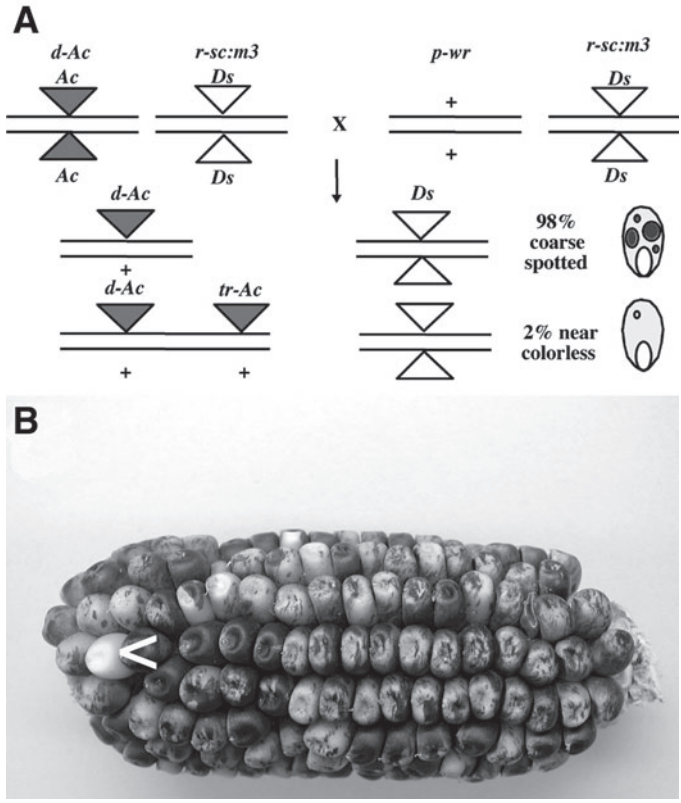


Fig. 3. Selection of *Ac* transposition events. **(A)** Crossing scheme. Lines homozygous for an *Ac* insertion are test-crossed to the *Ds* reporter line. New transposition events are detected as finely spotted kernels, as discussed in the text. **(B)** Typical test-cross ear derived from the cross shown in panel **A**. The majority of kernels are coarsely spotted, but one kernel (arrowhead) can be seen that conditions a near-colorless aleurone. Close examination of the aleurone and scutellum of this class of kernels will often reveal a very fine spotting pattern indicative of increased *Ac* copy number.

3. F1 plants are grown to maturity and can be screened for dominant mutant phenotypes. As all plants are maintained in a uniform W22 inbred background, subtle variations in plant stature, height, and flowering time can be readily identified.
4. F1 plants are self-pollinated.
5. Phenotypic screens are performed on segregating F2 families. Visual screens of F2 ears can be performed to identify embryo or endosperm-defective phenotypes. Alternatively, 20–25 kernels can be planted from each F2 ear for sandbench or field screens.
6. Co-segregation analysis using *Ac*-specific probes in DNA blot analysis can then

be performed to identify genomic fragments that carry an *Ac* insertion in a gene of interest ***(29,30)*** (*see* **Note 9**).

3.1.2. Directed Mutagenesis

1. A line(s) homozygous for an *Ac* insertion closely linked to gene of interest (<10 cM) should be test-crossed as females by plants homozygous for the mutation of interest (reference allele). Again, the reference allele must carry a recessive *r* allele (e.g., *r-g* or *r-r*) and preferable the *r-sc:m3* allele.
2. New transposition events are selected as finely spotted kernels as detailed above.
3. Phenotypic screens of F1 plants can be performed in greenhouse or in field screens to identify heteroallelic mutations consisting of an *Ac* insertion in the gene of interest and the reference allele.
4. Outcrosses or self-pollination of any heteroallelic mutants identified must be performed to ensure the recovery of putative *Ac*-induced mutations (*see* **Note 10**).
5. Backcrosses to the reference line and self-pollination of the recovered mutant individuals is expected to generate mostly mutant progeny with a few exceptional wild-type plants that may represent *Ac* revertant alleles.
6. All test-cross progeny to the W22 parent should be wild-type if the putative *Ac*-induced mutation segregates as a single recessive allele. Self-pollination of these test-cossed individuals should result in the segregation of either the reference allele or the putative *Ac*-induced allele.
7. Co-segregation analysis can then be performed using *Ac*-specific gene fragments as detailed in **Subheading 3.2.1.** (*see* **Note 11**).

3.2. Cloning Ac-Flanking Sequences

Traditionally, *Ac*-flanking sequences have been isolated in gene tagging experiments through the construction of genomic or subgenomic phage libraries. However, the construction of genomic libraries is both expensive and time-consuming. As an alternative, we have developed a PCR-based method to clone *Ac*-flanking insertions utilizing the IPCR technique ***(31)***.

3.2.1. DNA Blot Analysis

DNA blot analysis is first performed on segregating families to identify an *Ac*-containing restriction fragment-length polymorphism (RFLP) that co-segregates with the mutant phenotype using *Eco*RI or *Pst*I to fractionate DNA (*see* **Note 12**).

1. Extract total DNA from approx 1 g leaf tissue.
2. Digest DNA as follows:
   a. 3 μg Genomic DNA.
   b. 2 μL 10× Restriction enzyme buffer (Promega).
   c. 0.2 μL Bovine serum albumin (10 mg/mL).
   d. 0.2 μL RNase (10 mg/mL).

e. 10 U Restriction enzyme (Promega).
f. dH_2O to 20 μL.
g. Incubate at 37°C water bath for 3 h.

3. Fractionate DNA on 0.8% agarose gel overnight at low voltage (approx 30 V) with 5 μL DIG-labeled ladder (DNA Ladder VII, Roche Molecular Biochemicals).
4. Transfer DNA to Hybond-N^+ nylon membrane.
5. Synthesize DIG-labeled DNA probes using the PCR DIG Probe Synthesis kit according to manufacturer's recommendations.
6. Prehybridize blots for 30 min to 1 h at 43°C in 25 mL DIG Easy Hyb Buffer.
7. Dilute 3 μL probe (per blot) in 50 μL water, place in 100°C heat block for 5 min, and snap-cool on ice for 5 min.
8. Decant prehybridization buffer, add diluted probe to 5 mL Dig Easy Hyb, and add to hybridization tube.
9. Hybridize overnight at 43°C.
10. Remove probe solution and wash twice in low stringency wash at room temperature for 5 min.
11. Remove low stringency wash and wash twice in high stringency wash at 60°C for 15 min.
12. In Pyrex® pan, add enough maleic acid buffer to cover blots and shake at room temperature for 2 min.
13. Pour off maleic acid buffer and add enough blocking buffer to cover generously; shake at room temperature for at least 1 h (ideally 2 h).
14. Prepare diluted Anti-DIG-AP Fab fragments reagent (Roche Molecular Biochemicals). Spin the antibody for 5 min at 10,000*g*. Add 5 μL of DIG antibody to 50 mL blocking buffer per blot, and mix well. Replace blocking buffer with Dig antibody solution. Shake at room temperature for 30 min.
15. Pour off antibody solution and add washing buffer. Shake at room temperature for 15 min and repeat once.
16. Replace washing buffer with detection buffer and shake at room temperature for 3 min.
17. Dilute 20 μL CDP-*Star*™ reagent (Roche) in 2 mL detection buffer per blot. Place blots in plastic sheet protector (DNA side up) and squeeze out excess liquid. Pipet 2 mL of CDP-*Star* solution onto each blot. Let sit for 5 min at room temperature.
18. Remove blots from sheet protector, blot dry on Whatman paper and carefully wrap in Saran® Wrap.
19. Expose to X-ray film for 30 min or image on chemiluminescence detection system, such as Kodak® Image Station 440 CR (Eastman Kodak).
20. Blots can be stripped according to manufacturer's recommendations (DIG Easy Hyb) and reprobed using same protocol or stored at 4°C for future probing.
21. Identify an *Ac*-containing restriction fragment that contains <2.5 kb of flanking DNA. This equates to a fragment size of <7.1 kb for a *Pst*I digest (2.5 kb flanking plus 4.6 kb *Ac*) or <4.5 or 5.0 kb for *Eco*RI digests (2.5 kb flanking plus 2.0 or 2.5 kb *Ac*).

3.2.2. IPCR Protocol

1. Once a fragment of an appropriate size is identified, genomic DNA (approx 15 μg) is digested:
   a. 15 μg genomic DNA (0.3 μg/μL).
   b. 20 μL 10× Restriction enzyme buffer.
   c. 2 μL Bovine serum albumin (10 mg/mL).
   d. 2 μL RNase A (10 mg/mL).
   e. 50 U Restriction enzyme.
   f. dH_2O to 200 μL.
   g. Incubate at 37°C for 5 h.
2. Fractionate DNA on 0.8% agarose gel overnight at low voltage (*see* **Note 13**).
3. Under UV illumination, quickly cut above and below the region of interest with a scalpel. It is important to limit UV exposure to prevent DNA damage (*see* **Note 14**).
4. DNA is isolated using the GeneClean kit according to manufacturer's recommendations, with some modifications (*see* **Note 15** and **16**).
5. DNA is eluted in 2× 20 μL dH_20 and self-ligated as follows:
   a. 20 ng DNA fragment.
   b. 5 μL 10× Ligase buffer.
   c. 1 μL T4 DNA ligase (3 U/μL).
   d. dH_2O to 50 μL Final vol.
   e. The ligation is performed overnight at 4°C.
6. Add 50 μL Tris-EDTA (TE) to ligation and heat-kill at 65°C for 15 min.
7. Remove salts with Nucleotide Removal kit according to manufacturer's recommendations and elute in 50 μL TE.
8. Add 450 μL dH_2O and proceed to PCR.
9. To amplify the *Ac*-flanking regions, two rounds of PCR are performed using nested sets of *Ac*-specific primers. The *Ac* primer pairs used in the PCRs are listed in **Table 1** (*see* **Note 17**). First round PCR:
   a. 10 μL purified ligation products.
   b. 5 μL 10× Buffer containing 15 m*M* $MgCl_2$.
   c. 2 μL DMSO.
   d. 5 μL dNTP's (2 m*M*).
   e. 0.5 μL *Taq* DNA polymerase (5 U/μL).
   f. 2.5 μL Primer 1 (10 μ*M*)
   g. 2.5 μL Primer 2 (10 μ*M*).
   h. 22.5 μL dH_2O.
   PCR conditions: 94°C for 2 min for one cycle; 94°C for 30 min, 57°C for 30 s, 72°C for 1 min (1 min for every kb flanking sequence to amplify), cycle 30 times; 72°C for 10 min; hold at 4°C.
10. Dilute first round PCR product 1:200 in water and perform PCR as in first round but with nested primer pairs.
11. Fractionate 4 μL of PCR products to identify appropriately sized fragments.

12. If products are the predicted size (from the DNA blot), gel-purify the remaining PCR products using Gel Extraction Kit (*see* **Note 18**).
13. Resuspend DNA in 40 μL Elution Buffer (Qiagen).
14. Subclone fragments into pGEM-T Easy vector according to manufacturer's recommendations.

3.2.3. Verification of IPCR Products

As it is possible to amplify cryptic or somatic transposition events using this protocol, it is essential to verify the identity of recovered fragments.

1. Sequence analysis of the PCR product should reveal structurally intact *Ac* end sequences and provides the first indication that sequences flanking an active element have been isolated.
2. To confirm the identity of the cloned product, an *Ac*-flanking fragment should be labeled and used on a segregating family in DNA blot analysis (*see* **Note 19**).
3. A readily discernable size shift should be detectable in segregating mutant individuals when compared to wild-type parental lines.

4. Notes

1. Because of the greater genetic resolution afforded by the IBM population ***(32)***, we first try to position sequences immediately flanking *Ac* elements on the IBM map. If polymorphisms are not readily detected in the parental B73 and Mo17 lines, one of two Brookhaven National Laboratory (BNL) populations are utilized. Efforts are now underway to link marker data between multiple recombinant inbred lines and can be used to move between the IBM and BNL populations (http://www.agron.missouri.edu/cMapDB/cMap.html).
2. Each nontransgenic line is genotyped in the Brutnell laboratory by DNA blot analysis prior to distribution to examine both *Ac* copy number and *Ac* position within the genome.
3. Homozygous lines display subtle variations in aleurone spotting pattern detailed on the project Web site (http://bti.cornell.edu/Brutnell_lab2/Projects/Tagging/BMGG_pro_tagging.html). This variation in *Ds*-mediated variegation pattern most likely reflects positional effects on *Ac* expression ***(33)*** and can sometimes be used to distinguish one homozygous line from another (T. Brutnell, unpublished observations).
4. Deviations in the pattern of variegation most likely reflect increases or decreases in *Ac* copy number that will confound the selection of new transposition events. In addition, germinal transposition of the *Ds* reporter, will result in approx 10–20% fully colored kernels, preventing the monitoring of *Ac* in the genome. Thus, the pattern of aleurone variegation should be examined closely in each kernel selected to enrich for a population of seed that is homozygous for a single *Ac* insertion and the *Ds* reporter.
5. In DNA blot analysis, a fragment of a predicted size should be visible, and no

additional *tr-Acs* should be present in the families. PCR analysis is performed using the *Ac* and gene-specific primer pairs to confirm the precise location of the element in the genome (details found at http://bti.cornell.edu/Brutnell_lab2/Projects/Tagging/BMGG_pro_tagging.html). These assays should be performed on pooled progeny tissue. For instance, if 10 self-pollinated ears are generated for mutagenesis, approx 100 kernels/ear should display the expected pattern of *Ds*-mediated variegation as detailed above. If all of the seed from an ear is planted for use in mutagenesis, there will be 10 families of 100 individuals for a total of 1000 putative *tr-Ac/tr-Ac*, *r-sc:m3/r-sc:m3* plants. Prior to performing crosses, leaf samples should be taken from 10 randomly selected individuals/ear and pooled for DNA extraction. DNA blot analysis and PCR analysis should then be performed on the 10 DNA samples representative of 100 individuals in the population (e.g., **Fig. 2**).

6. As mentioned above, too few regional mutagenesis experiments have been performed to accurately predict the number of transposition events necessary for a successful tagging experiment. Nevertheless, based on a limited number of experiments in our laboratory and elsewhere ***(22,34)***, we estimate that 1000–3000 transpositions (F1 individuals) should be sufficient to recover an insertion allele if the gene is within 10 cM of a donor *Ac*. In most cases, this will equate to approx 500–1500 crosses, assuming a transposition frequency of 2% and the recovery of approx 100 k/ear.
7. Most kernels inherit a single donor *Ac* from the maternal tissues, resulting in a characteristic coarse spotted aleurone. However, approx 2–4% of the progeny will carry a transposed *Ac* (*tr-Ac*) in addition to the *d-Ac*. This increase in *Ac* copy number results in later *Ds* excisions from the *r* locus and, thus, smaller revertant sectors.
8. In some instances, the increase in *Ac* copy number results in a completely colorless aleurone. To ensure that the *d-Ac* and the *tr-Ac* are transmitted to the embryo, it is important to examine the variegation pattern of the scutellum tissue in addition to the endosperm. The scutellum is derived from diploid embryonic tissues and will, therefore, more accurately reflect the genotype of the mature plant. As a general rule, we screen for new transpositions as conditioning a finely spotted aleurone and colorless scutellum, a finely spotted scutellum and colorless aleurone, or a finely spotted aleurone and a finely spotted scutellum. In pilot studies where kernels with a colorless aleurone and colorless scutellum were selected, less than 50% of the resulting ears contained a *tr-Ac* (400 out of 823). Thus, evidence for some *Ac* activity in either the endosperm or scutellum of F1 seed will greatly enrich for kernels that carry new transposition events.
9. We have had the most success at identifying *Ac* fragments with the methylation insensitive restriction enzyme *Eco*RI. This enzyme cleaves in the middle of *Ac*, resulting in DNA fragments that contain 2.0 and 2.5 of *Ac* sequence. Blots are sequentially hybridized to 700 and 900 bp *Eco*RI-*Hin*dIII *Ac* internal fragments (e.g., ***30***). Although these fragments detect several of the immobile *Ac*-like or cryptic *Ac* sequences ***(35)***, active elements appear as novel bands not present in

either inbred parent (*see* **Fig. 2A**). Alternatively, the methylation-sensitive restriction enzyme *Pst*I can be used to identify *Ac*-containing RFLPs. The majority of cryptic *Ac*s are located in heavily methylated regions of the genome ***(36)***, but active *Ac*s often insert in hypomethylated regions of the genome ***(37)***. If the same active *Ac* is detected with multiple enzymes, it is easier to clone a fragment using a methylation-sensitive enzyme, such as *Pst*I, as most of the cryptic fragments will remain in the high molecular weight fraction of the gel.

10. Staggered plantings, using the inbred W22 parent and the reference allele line, should be planted 7–10 d before and after the F1 progeny screen is planted.
11. If a hybrid population is created (directed mutagenesis), methylation-sensitive restriction enzymes should be utilized to identify newly transposed *Ac* elements.
12. In practice, it is prudent to use several methylation-sensitive and -insensitive enzymes in co-segregation analysis to ensure that a fragment of an appropriate size can be isolated for use in the IPCR protocol. For example, we have detected 75% of active *Ac* elements with *Eco*RI, but only 57% were within the appropriate size range for IPCR. In our experience, it has been difficult to amplify target sequences greater than 2.5 kb.
13. Tape the lanes of a comb together to load approx 200 µL of DNA digest and include DNA marker lanes on either side.
14. Excise the thinnest possible gel slice in order to get good yield from the Gene Clean III kit. The gel slice should weigh ≤1 g and no more than 1.5 g. For example, if the *Ac*-containing fragment is 3 kb, then cut just above and below the 3 kb mark to ensure that the *Ac* fragment is included in the gel slice.
15. In GeneClean protocol, use 20 µL of glass milk, regardless of the vol of NaI solution, and allow glass milk annealing to occur for 30 min, rather than the manufacturer's recommended 5 min.
16. Prior to IPCR amplification, ensure that appropriately sized fragment(s) were isolated through DNA blot analysis. Probe with the same *Eco*RI-*Hin*dIII *Ac* internal fragments used to detect the active *Ac*.
17. If additional PCR primers are designed, make sure they are not >200 bp from the end of the *Ac* sequence. This will add unnecessary sequence and lengthen your products, thus reducing the efficiency of PCR. Primers should be between 20–30 bases, 50% GC-rich, and have a melting temperature (T_m) ≥ 60°C.
18. If PCR products are not detected after the second round of PCR, a third round can be performed with another set of nested primers designed to the specification in **Note 17**. Dilute the second round product 1:200 in water and use the same cycling conditions. Multiple bands may result from third round amplification. If so, the predicted size band should be most intense.
19. In most instances, sequences immediately adjacent to an active *Ac* element are represented as single or low copy elements in the genome ***(26)***. Thus, PCR primers should be designed as close to the ends of *Ac* as possible to generate a gene-specific fragment as probes.

References

1. McClintock, B. (1947) Cytogenetic studies of maize and *Neurospora*. *Carnegie Institution of Washington Year Book* **46,** 146–152.
2. Kunze, R., Saedler, H., and Lönnig, W.-E. (1997) Plant transposable elements, in *Advances in Botanical Research, Vol. 27*. (Callow, J. A., ed.), Academic Press, London, pp. 332–469.
3. Bancroft, I., Bhatt, A. M., Sjodin, C., Scofield, S., Jones, J. D., and Dean, C. (1992) Development of an efficient two-element transposon tagging system in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **233,** 449–461.
4. Bhatt, A. M., Page, T., Lawson, E. J., Lister, C., and Dean, C. (1996) Use of *Ac* as an insertional mutagen in *Arabidopsis*. *Plant J.* **9,** 935–945.
5. Fedoroff, N. V. and Smith, D. L. (1993) A versatile system for detecting transposition in *Arabidopsis*. *Plant J.* **3,** 273–289.
6. Sundaresan, V., Springer, P., Volpe, T., et al. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9,** 1797–1810.
7. Martienssen, R. A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl. Acad. Sci. USA* **95,** 2021–2026.
8. Walbot, V. (1992) Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. *Annu. Rev. Plant. Physiol. Plant Mol. Biol.* **43,** 49–82.
9. Brutnell, T. P. and Dellaporta, S. L. (1994) Somatic inactivation and reactivation of *Ac* associated with changes in cytosine methylation and transposase expression. *Genetics* **138,** 213–225.
10. Van Schaik, N. W. and Brink, R. A. (1959) Transposition of *Modulator*, a component of the variegated pericarp allele in maize. *Genetics* **44,** 725–738.
11. Greenblatt, I. M. (1984) A chromosome replication pattern deduced from pericarp phenotypes resulting from movements of the transposable element, *Modulator*, in maize. *Genetics* **108,** 471–485.
12. Dooner, H. K. and Belachew, A. (1989) Transposition pattern of the maize element *Ac* from the *bz-m2*(*Ac*) allele. *Genetics* **122,** 447–457.
13. Brutnell, T. P. (2002) Transposon tagging in maize. *Funct. Integr. Genomics* **2,** 4–12.
14. Talbert, L. E., Patterson, G. I., and Chandler, V. (1989) *Mu* transposable elements are structurally diverse and distributed throughout the genus *Zea*. *J. Mol. Evol.* **29,** 28–39.
15. Orton, E. R. and Brink, R. A. (1966) Reconstitution of variegated pericarp allele in maize by transposition of *Modulator* back to *P* locus. *Genetics* **53,** 7–16.
16. Peterson, T. (1990) Intragenic transposition of Ac generates a new allele of the maize *P* gene. *Genetics* **126,** 469–476.
17. Moreno, M. A., Chen, J., Greenblatt, I., and Dellaporta, S. L. (1992) Reconstitutional mutagenesis of the maize *P* gene by short-range *Ac* transpositions. *Genetics* **131,** 939–956.

18. Athma, P., Grotewold, E., and Peterson, T. (1992) Insertional mutagenesis of the maize *P* gene by intragenic transposition of *Ac*. *Genetics* **131,** 199–209.
19. Schauser, L., Roussis, A., Stiller, J., and Stougaard, J. (1999) A plant regulator controlling development of symbiotic root nodules. *Nature* **402,** 191–195.
20. Dellaporta, S. L. and Moreno, M. A. (1994) Gene tagging with *Ac/Ds* elements in maize, in *The Maize Handbook* (Freeling, M. and Walbot, V., eds.), Springer, New York, pp. 219–233.
21. Colasanti, J., Yuan, Z., and Sundaresan, V. (1998) The indeterminate gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. *Cell* **93,** 593–603.
22. Dellaporta, S. L., Greenblatt, I. M., Kermicle, J. L., Hicks, J. B., and Wessler, S. R. (1988) Molecular cloning of the maize R-nj allele by transposon tagging with *Ac*, in *Chromosome Structure and Function: Impact of New Concepts* (Gustafson, J. P. and Appels, R., eds.), Plenum Press, New York, pp. 263–282.
23. McClintock, B. (1948) Mutable loci in maize. *Carnegie Institution of Washington Year Book* **47,** 155–169.
24. Auger, D. L. and Sheridan, W. F. (1994) Using cytogenetics to enhance transposon tagging with *Ac* throughout the maize genome, in *The Maize Handbook* (Freeling, M. and Walbot, V., eds.), Springer-Verlag, New York, pp. 234–239.
25. Auger, D. L. and Sheridan, W. (1999) Maize stocks modified to enhance the recovery of *Ac*-induced mutations. *J. Hered.* **90,** 453–459.
26. Cowperthwaite, M., Park, W., Xu, Z., Yan, X., Maurais, S. C., and Dooner, H. K. (2002) Use of the transposon *Ac* as a gene-searching engine in the maize genome. *Plant Cell* **14,** 713–726.
27. Dooner, H. K. and Kermicle, J. L. (1971) Structure of the *R-r* tandem duplication in maize. *Genetics* **67,** 437–454.
28. Kermicle, J. (1984) Recombination between components of a mutable gene system in maize. *Genetics* **107,** 489–500.
29. Schultes, N. P., Brutnell, T. P., Allen, A., Dellaporta, S. L., Nelson, T., and Chen, J. (1996) *Leaf permease1* gene of maize is required for chloroplast development. *Plant Cell* **8,** 463–475.
30. Schultes, N. P., Sawers, R. J. H., Brutnell, T. P., and Krueger, R. W. (2000) Maize *high chlorophyll fluorescent 60* mutation is caused by an *Ac* disruption of the gene encoding the chloroplast ribosomal small subunit protein 17. *Plant J.* **21,** 317–327.
31. Ochman, H., Gerber, A. S., and Hartl, D. L. (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics* **120,** 621–623.
32. Lee, M., Sharopova, N., Beavis, W. D., et al. (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.* **48,** 453–461.
33. Brutnell, T. P., May, B. P., and Dellaporta, S. L. (1997) The *Ac-st2* element of maize exhibits a positive dosage effect and epigenetic regulation. *Genetics* **147,** 823–834.
34. DeLong, A., Calderon-Urrea, A., and Dellaporta, S. L. (1993) Sex determination

gene *TASSELSEED2* of maize encodes a short-chain alcohol dehydrogenase required for stage-specific floral organ abortion. *Cell* **74,** 757–768.

35. Leu, J. Y., Sun, Y. H., Lai, Y. K., and Chen, J. (1992) A maize cryptic *Ac*-homologous sequence derived from an *Activator* transposable element does not transpose. *Mol. Gen. Genet.* **233,** 411–418.
36. Chen, J., Greenblatt, I. M., and Dellaporta, S. L. (1987) Transposition of *Ac* from the *P* locus of maize into unreplicated chromosomal sites. *Genetics* **117,** 109–116.
37. Chomet, P. S., Wessler, S., and Dellaporta, S. L. (1987) Inactivation of the maize transposable element *Activator* (*Ac*) is associated with its DNA modification. *EMBO J.* **6,** 295–302.

11

T-DNA Mutagenesis in *Arabidopsis*

Jose M. Alonso and Anna N. Stepanova

Summary

Insertional mutagenesis is a basic genetic tool that allows for a rapid identification of the tagged genes responsible for a particular phenotype. Transposon and *Agrobacterium*-mediated DNA integration are the most commonly used biological mutagens in plants. The main drawback of these technologies is the relatively low frequency of mutations, as compared to those induced by conventional chemical or physical agents, thus limiting the use of insertional mutagens to the generation of large mutant populations in few genetic backgrounds. Recent improvements in *Agrobacterium*-mediated transformation efficiency and an increasing repertoire of transformation vectors available to the research community is making this type of mutagen very attractive for individual laboratories interested in the studies of mutations in particular genetic backgrounds. Herein, we describe a simple yet robust *Arabidopsis* transformation procedure that can be used to generate large numbers of insertional mutants in *Arabidopsis thaliana*. Using this protocol, transformation efficiencies of up to 5% can be achieved.

Key Words

Arabidopsis, plant transformation, *Agrobacterium*, mutagenesis, T-DNA, vacuum infiltration

1. Introduction

Agrobacterium-mediated transformation has been used extensively to incorporate selected DNA sequences into the genome of the target cells. Although *Agrobacterium* can efficiently transfer bits of its DNA to a number of organisms including human cells ***(1)***, its natural target is the plant genome ***(2)***. The mechanisms involved in the transfer and integration of the foreign DNA into the plant genome are currently under investigation, and several bacterial and, more recently, plant molecular components that participate in this process have been determined ***(3)***.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

In nature, *Agrobacterium tumefaciens* infects the roots of wounded plants and transfers a fragment of its tumor-inducing (Ti) plasmid, the so-called transferred DNA (T-DNA), into the plant genome ***(4)***. The T-DNA is delimited by two regions called left and right border (LB and RB, respectively) that are comprised of 25-bp-long repeats. Interestingly, the naturally occurring bacterial sequences that reside between the LB and RB on the Ti plasmid are not essential for the transfer and integration of the T-DNA and are, therefore, dispensable. One can artificially replace these sequences with a fragment of foreign DNA and, thus, take advantage of the natural capacity of *Agrobacterium* and target this DNA of interest, along with the border sequences, into the genome of the host cell ***(5)***. A large number of vectors originally derived from the basic Ti plasmid have been constructed. A common feature of these vectors is the presence in the T-DNA of a selectable marker (a gene that confers resistance to a particular drug), which allows for a positive selection of transformants. In addition to the marker, other sequences may be introduced into the T-DNA portion of the Ti plasmid-derived vectors. Based on the T-DNA composition, three main categories of vectors can be distinguished. Transfer vectors are probably the most widely used in modern laboratories. They are designed to insert the sequences of interest (for example, a fragment of your favorite gene) into the plant genome. Common applications of this class of vectors include overexpression or antisense studies, complementation analyses, reporter fusions, etc. Most of the transfer vectors can also be used to generate simple knock-outs. For this purpose, only the selectable marker gene and transcription terminators are kept between the left and right T-DNA borders. The second class of vectors is comprised of promoter-trap, enhancer-trap, and gene-trap vectors used to generate transgenic plants in which a reporter gene adopts a specific pattern of expression dependent on the region of the genome where the T-DNA is integrated. The expression patterns of the reporter genes typically reflect the activity of the tagged genes ***(6)***. Therefore, these vectors are commonly employed to generate and identify mutants with the desired patterns of spatial–temporal reporter expression and, ideally, clone the respective genes. The third class is represented by activation tagging vectors designed to boost the expression of the gene or genes in the vicinity of the integration site. This is achieved by incorporating tandems of transcriptional enhancers close to one of the T-DNA borders ***(7)***. All three kinds of vectors can be used in random mutagenesis, and the selection of a particular type of vector over the others depends on the specific goals of the mutagenesis (i.e., gain- or loss-of-function screening) (*see* **Table 1**, and Chapters 20 and 21 in this book).

Agrobacterium can transform a number of plant species, but this chapter will focus on the transformation of the reference plant *Arabidopsis thaliana*. Fifteen years ago, the transfer of a DNA sequence into the *Arabidopsis* genome

Table 1
Main Characteristics of the T-DNA Mutagenesis Systems used in *Arabidopsis*

Plasmid	Bacterial selection	Plant selection	Features	*Agrobacterium* strain	Method of transformation	Reference
pGKB5	Kanamycin	Basta	Enhancer trap	C58C1(pMP90)	Vacuum-infiltration	***(11)***
pD991	Gentamycin	Kanamycin	Enhancer trap	Not reported	Vacuum-infiltration	***(19)***
pSKI015	Carbenicillin	Basta	Activation tagging	GV3101 (pMP90RK)	Vacuum-infiltration	***(7)***
pSKI074	Carbenicillin	Kanamycin	Activation tagging	GV3101 (pMP90RK)	Vacuum-infiltration	***(7)***
pYU565	Spectinomycin	Basta	Knock-out	GV3101 (pMP90RK)	Floral dip	***(20)***
35SpBARN	Kanamycin	Basta	cDNA overexpression	GV3101 (pMP90RK)	Floral dip	***(21)***
pMOG553	Kanamycin	Hygromycin	Promoter trap	MOG101 (C58 derivative)	Root explants	***(22)***
pPCV6NFHyg	Ampicillin	Hygromycin	Translational fusion	Not reported	Stem, leaf, and root explants	***(23)***
PPCV621	Ampicillin	Hygromycin	Transcriptional fusion	Not reported	Stem, leaf, and root explants	***(23)***

required a long, labor-intensive, yet low-efficiency procedure. Bacterial culture was co-incubated with root or leaf explants. Upon infection, the callus was obtained from the transformed plant cells and induced to form shoots and roots to regenerate whole plants ***(8)***. The process required skilled personnel and adequate equipment for plant tissue manipulation, thus limiting its success to only few laboratories. In the past decade, seed and, more recently, the "*in planta*" procedures have simplified enormously the transformation protocol, making it possible to generate large numbers of transformants with minimum requirements in personnel and laboratory equipment. Furthermore, these new methods eliminated the problem of undesired somaclonal mutations that frequently appeared during plant tissue culture growth and regeneration ***(9,10)***. Among the *in planta* procedures, the floral dip and vacuum infiltration currently are the most popular methods of plant transformation and are described in detail in this chapter ***(11,12)***. The basic idea behind these procedures is to bring the *Agrobacterium* culture in close contact with those plant cells that are prone to transformation. A number of studies have been conducted to determine which plant cells can and cannot be transformed and what developmental stages are the most susceptible. An important clue to answer this question came from the observation that all the primary transformed plants (T1) are hemizygous, suggesting that the transfer occurred after the divergence between the anther and ovary cell lineages ***(9,13)***. Recent studies elegantly demonstrated that the ovule and, more specifically, the female gametophyte chromosomes are the main targets of the T-DNA integration ***(14,15)***. A very important practical implication of these findings is that each one of the primary transformant plants obtained is the result of an independent T-DNA integration event and therefore represents an independent insertional mutant.

2. Materials

2.1. Plant Growth

1. Metromix-200 soil or equivalent.
2. Germination trays (21 × 11 × 1 1/4 inch; Hummert International; cat. no. F1221).
3. Propagation dome (Hummert International; cat. no. CW221).
4. Square plastic pots (3 1/2 inches; Hummert International; cat. no. KT1835).
5. 500-mL Squibb separatory funnel.

2.2. Agrobacterium Culture

1. Flasks.
2. Temperature controlled shaker.
3. Rifampicin (Sigma).
4. Kanamycin monosulfate (FisherBiotech®).

5. Petri dishes.
6. Agar.

2.3. Plant Infiltration

1. Elastic rubber bands.
2. Vacuum pump.
3. Vacuum chamber.
4. Removable-cover pipet tip racks (Rainin Instruments, cat. no. RT-L1000).
5. Infiltration media: 5% sucrose, one-half strength Murashige & Skoog salts (optional), 44 n*M* benzylamino purine (10 µL/L of a 1 mg/mL stock in dimethylsulfoxide [DMSO]) (optional). Adjust pH to 6.0 with 1 *M* KOH and then add Silwet L-77 to 0.02%.
6. Benzylamino purine (Sigma).
7. Silwet L-77 (Lehle Seeds).
8. J2-HC centrifuge or equivalent (Beckman Coulter).
9. JA-10 rotor or equivalent (Beckman Coulter).
10. 500 mL centrifuge bottles.

2.4. Selection of Transformants

1. Murashige & Skoog salt mixture (Life Technologies).
2. Hygromycin B (Life Technologies).
3. Basta (Finale) (AgrEvo).
4. Phosphinothricin (PPT) (glufosinate ammonium) (Crescent Chemical).
5. HCl.
6. Bleach.
7. Petri dishes.
8. Agar.
9. Disposable 50-mL centrifuge tubes.
10. Fine-pointed forceps.
11. Balances.

3. Methods

3.1. Plant Growth

1. Combine seeds (200 seeds per 21 × 11 inch tray, which is equivalent to approx 10 seeds per 3 1/2 inch square pot) of the desired genetic background with 700 µL of distilled water in a 1.5-mL microcentrifuge tube (*see* **Note 1**).
2. Stratify the seeds for 3 d at 4°C.
3. Add seeds to 200 mL of 0.1% agar in water that had been previously melted and cooled to room temperature, then mix to uniformity.
4. Transfer the mixture to a Squibb separatory funnel and disperse the seeds uniformly on the soil surface.
5. Grow the plants at a constant temperature of 21°C in a light–dark cycle of 16 to 8 h. Keep trays of plants covered with transparent plastic domes for the first 2 wk

to maintain high humidity (*see* **Note 2**). Approximately 4 to 5 wk after planting (this period may change significantly depending on the genetic background of the plants) floral stems reach about 15 cm in length and are then ready to be infiltrated (*see* **Note 3**).

3.2. Agrobacterium *Culture*

1. In a 250-mL flask, inoculate a single colony of *Agrobacterium* cells (e.g., C58C1 pMP90) ***(16)*** that harbor the desired vector into 50 mL of LB media supplemented with the appropriate antibiotic plus 25 μg/mL rifampicin (*see* **Note 4**).
2. Grow cells for 12–24 h at 30°C with constant shaking at 120 rpm.
3. Use this culture (entire 50 mL) to inoculate 1 L of fresh LB media supplemented with the appropriate antibiotic in a 2-L flask.
4. Grow the culture for 12–16 h at 30°C with constant shaking at 120 rpm.
5. Concentrate *Agrobacterium* cells by centrifugation at 5000 rpm for 10 min in a JA-10 rotor (or its equivalent) at room temperature.
6. Discard the supernatant (*see* **Note 5**) and resuspend the cells corresponding to 500 mL of culture in an equal vol of infiltration media (*see* **Note 6**).

*3.3. Plant Infiltration (*see *Note 7)*

1. Transfer the *Agrobacterium* suspension (the infiltration mixture) to an appropriate container (e.g., a plastic tip rack can be used) (*see* **Note 8**).
2. Keep the soil from falling out of the pot with the help of two rubber bands (*see* **Note 9**).
3. Invert the pot with plants up-side-down and submerge all inflorescences into the infiltration mixture.
4. Transfer the plants while still submerged in the infiltration mixture into the vacuum chamber.
5. Apply vacuum by turning on a diaphragm pump for about 30–60 s or until the infiltration mixture starts bubbling.
6. Rapidly release the vacuum by disconnecting the infiltration chamber from the pump.
7. Repeat (optional) the infiltration step by reconnecting the chamber to the pump and, again, break the vacuum after 30–60 s.
8. Immediately transfer the infiltrated plants to a new tray placing the pots horizontally.
9. Cover the trays with a transparent dome and allow plants to recover in the growth chamber.
10. Twenty-four hours after transformation, gradually increase aeration in the tray (*see* **Note 2**).
11. Two days after the infiltration uncover the plants completely, put them in a vertical position and water generously.
12. Allow plants to set seeds, water them as needed for 2–4 wk.
13. Dry the plants by stopping watering.

3.4. Selection of Transformants

3.4.1. Selection of Transformants in Soil (Basta/Finale Selection)

1. Collect and clean the primary transformation (T1) seeds by passing them through a plastic mesh or a sieve.
2. Uniformly distribute stratified seeds (*see* **Note 10**) on the surface of wet soil as described in **Subheading 3.1.**, **steps 1–4**.
3. Grow the plants at a constant temperature of 21°C in a light–dark cycle of 16 to 8 h. Keep trays of plants covered with transparent plastic domes for the first 2 wk to maintain high humidity (*see* **Note 2**).
4. Spray 2-wk-old plants with a 0.017% aqueous solution of Finale and repeat application 2 wk later (*see* **Note 11**).
5. Grow the surviving plants to maturity at 21°C in a light–dark cycle of 16 to 8 h watering as needed.
6. Harvest seeds from individual T1 plants or small pools of T1s (*see* **Note 12**).
7. Perform phenotypic screening in the next (T2) generation (*see* **Note 13**).

3.4.2. Selection of Transformants in Plates (Kanamycin, Hygromycin B, or PPT Selection)

1. Aliquot up to 20,000 cleaned T1 seeds into a 50-mL disposable centrifuge tube and place the tube uncovered in a 5-L dessicator chamber. Perform manipulations described in **steps 2–4** in a fume hood.
2. Place a 250-mL flask containing 100 mL bleach inside of the dessicator chamber.
3. Add 4 mL of concentrated HCl to the flask and immediately seal the chamber (*see* **Note 14**).
4. Expose the seeds to the chlorine gas atmosphere for 1 h (*see* **Note 15**).
5. Transfer the centrifuge tube with the seeds to a sterile laminar hood and let the chlorine vapors dissipate for at least 12 h.
6. Spread the dry sterilized seeds on the surface of MS plates supplemented with the appropriate selection drug (*see* **Note 16**).
7. Stratify the seeds by incubating the plates for 3 d at 4°C.
8. In a sterile laminar hood, shake off excessive condensation from the plate lids.
9. Allow the seeds to germinate for 3 d in the dark at 21°C.
10. Move plates to constant light and grow seedlings at 21°C for an additional 7–10 d (*see* **Note 17**).
11. Transfer the surviving plants to wet soil using fine-pointed forceps.
12. Cover trays with plastic domes for the first 2 wk of growth (*see* **Note 2**).
13. Grow plants to maturity at 21°C in a light–dark cycle of 16 to 8 h watering as needed.
14. Harvest seeds from individual T1 plants or small pools of T1s (*see* **Note 12**).
15. Perform phenotypic screening in the next (T2) generation (*see* **Note 12**).

4. Notes

1. One thousand *Arabidopsis* seeds weigh about 20 mg. Use this estimate to weigh out the appropriate number of seeds. Transformation of 400 plants usually yields 1000–5000 individual T-DNA lines.
2. To avoid abrupt changes in humidity, aeration in the tray is increased gradually by shifting the plastic domes approx 1 inch with respect to the tray 24 h prior to removing them completely.
3. Although an increase in the transformation efficiency has been reported when the first appearing floral stems were removed and new inflorescences were allowed to form prior to performing the transformation procedure ***(12)***, in our hands this amendment to the protocol resulted in the reduction of the transformation efficiency, presumably, by decreasing the vigor of plants.
4. Water-soluble antibiotics solutions should be filter-sterilized. Rifampicin (diluted from 25 mg/mL stock in methanol) allows for selection of slower growing C58C1 pMP90 against potential bacterial contamination. All antibiotics stock solutions should be stored at –20°C.
5. All materials that came in contact with *Agrobacterium* should be sterilized.
6. Resuspended cells should be used within the next few hours.
7. Even though more reproducible results are obtained with the vacuum infiltration protocol, floral dipping is a simpler and faster procedure. **Steps 4** through **7** from **Subheading 3.3.** should be skipped if the latter method is used. Watering plants 1 d prior to transformation may improve the efficiency and reproducibility of the dipping protocol.
8. Enough media should be added to the tip container to be able to submerge all of the inflorescences, but not to contaminate the soil.
9. To avoid soil falling out of the pot, two rubber bands are placed around the pot (along the vertical axes of the pot) and under the leaves of the plants holding together the pot and the soil. In our hands, placing rubber bands immediately prior to transformation works better than growing the plants through a fabric mesh, with the latter method often resulting in fungal or algal contamination.
10. Optimal seed density depends on the efficiency of transformation. While the efficiency somewhat varies dependent on the vector–plant genetic background combination, the average efficiency achieved with this protocol is 1 to 2%. Therefore, planting of at least 5000 T1 seeds per 21 × 11 × 1 1/4 inch tray is recommended (*see* **Note 1**).
11. Spraying should be performed in ventilated areas. Gloves and face shield or safety glasses should be used when handling this product. See Material Safety data sheet (provided with the product) for more information.
12. Propagation of individual lines or small pools is highly recommended when sterility or lethality are expected in the transformed plants.
13. In some cases (e.g., when dominant mutations are expected, such as in the activation tagged lines), phenotypic screening of adult plants can be performed as early as in the T1 generation. About two-thirds of all mutations do not co-segregate with the resistance marker ***(17)***, and therefore, it is recommended that segregation

analysis be performed prior to identification of the T-DNA insertion sites. Once the linkage between the T-DNA and the phenotype has been established, the insertion sites can then be recovered using plasmid rescue, inverse polymerase chain reaction (PCR) or thermal asymmetric interlaced (TAIL)-PCR approaches (*see* Chapter 15 by Tatjana Singer in this book). Insertions in the genes of interest can also be identified by extracting DNA from the transformed plants and screening by PCR for the presence of a T-DNA insertion in the proximity of the desired gene ***(18)***. Identification of several alleles, complementation or recapitulation experiments should be performed to conclude that the tagged gene identified is, in fact, responsible for the mutant phenotype.

14. When HCl reacts with the bleach, chlorine fumes are immediately released, and small drops of bleach/HCl may jump onto the seeds. This should be avoided by mixing bleach and HCl in a flask with long thin neck rather than a beaker. Chlorine fumes are very toxic and corrosive and should be handled with extreme care.
15. Although 1 h works well in most of the cases, different seed batches may behave very differently. Too little time will result in incomplete sterilization and contamination problems, whereas longer sterilization may kill the seeds. Therefore, it is recommended that a small aliquot of seeds initially be tested before all of the seeds are sterilized.
16. To reduce contamination problems, MS plates without sugar may be used. The following drug concentrations are typically used for the selection of transformed plants in plates: 50–100 μg/mL kanamycin, 10–100 μg/mL hygromycin B, or 20–50 μ*M* PPT (the active ingredient of Basta). The recommended seed number is 3000 per 15 × 150 mm Petri plate (*see* **Note 1**).
17. Seal the plates with parafilm to avoid excessive water loss. Transformed plants can often be identified as early as a few hours after the plates were moved to constant light. After 1 to 2 d in the light kanamycin- and PPT-resistant plants show dark green cotyledons, whereas sensitive plants remain bleached. Conversely, upon exposure to light, hygromycin-sensitive plants "green" normally, but can be distinguished from the resistant seedlings after an additional 3–7 d by the retarded growth of both roots and hypocotyls. Importantly, root elongation of kanamycin-, PPT-, and hygromycin-sensitive plants is greatly inhibited by all three substances, whereas roots of resistant plants grow normally and exceed 5 mm in length 5–7 d postgermination.

References

1. Kunik, T., Tzfira, T., Kapulnik, Y., Gafni, Y., Dingwall, C., and Citovsky, V. (2001) Genetic transformation of HeLa cells by *Agrobacterium. Proc. Natl. Acad. Sci. USA* **98,** 1871–1876.
2. Hooykaas, P. J. J. and Beijersbergen, A. G. M. (1994) The virulence system of *Agrobacterium tumefaciens*. *Ann. Rev. Phytopathol*. **32,** 157–179.
3. Tzfira, T. and Citovsky, V. (2002) Partners-in-infection: host proteins involved in the transformation of plant cells by *Agrobacterium. Trends Cell Biol*. **12,** 121–129.

4. Gelvin, S. B. (1998) The introduction and expression of transgenes in plants. *Curr. Opin. Biotechnol*. **9,** 227–232.
5. Zambryski, P. (1992) Chronicles from the *Agrobacterium*-plant cell DNA transfer story. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43,** 465–490.
6. Springer, P. S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell*. **12,** 1007–1020.
7. Weigel, D., Ahn, J. H., Blazquez, M. A., et al. (2000) Activation tagging in *Arabidopsis. Plant Physiol*. **122,** 1003–1013.
8. Valvekens, D., Montague, M. V., and Lijsbettens, M. V. (1988) *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. *Proc. Natl Acad. Sci. USA* **85,** 5536–5540.
9. Bechtold, N., Ellis, J., and Pelletier, G. (1993) *In planta Agobacterium*-mediated gene transfer by infiltration of adult *Arabidopsis* plants. *C R Acad. Sci. Paris Life Sci*. **316,** 1194–1199.
10. Feldmann, K. A. and Marks, M. D. (1987) *Agrobacterium*-mediated transformation of germinating seeds of *Arabidopsis thaliana*: a non-tissue culture approach. *Mol. Gen. Genet.* **245,** 704–715.
11. Bechtold, N. and Pelletier, G. (1998) *In planta Agrobacterium*-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. *Methods Mol. Biol*. **82,** 259–266.
12. Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana. Plant J*. **16,** 735–743.
13. Feldmann, K. (1992) T-DNA insertion mutagenesis in *Arabidopsis*: seed infection/transformation, in *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H., and Schell, J., eds.), World Scientific Publishing, Singapore, pp. 274–289.
14. Bechtold, N., Jaudeau, B., Jolivet, S., et al. (2000) The maternal chromosome set is the target of the T-DNA in the *in planta* transformation of *Arabidopsis thaliana. Genetics* **155,** 1875–1887.
15. Ye, G. N., Stone, D., Pang, S. Z., Creely, W., Gonzalez, K., and Hinchee, M. (1999) *Arabidopsis* ovule is the target for *Agrobacterium in planta* vacuum infiltration transformation. *Plant J*. **19,** 249–257.
16. Koncz, C. and Schell, J. (1986) The promoter of the Tl-TDNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Mol. Gen. Genet.* **204,** 383–396.
17. McElver, J., Tzafrir, I., Aux, G., et al. (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana. Genetics* **159,** 1751–1763.
18. Winkler, R. G., Frank, M. R., Galbraith, D. W., Feyereisen R., and Feldmann, K. A. (1998) Systematic reverse genetics of transfer-DNA-tagged lines of *Arabidopsis*. Isolation of mutations in the cytochrome P450 gene superfamily. *Plant Physiol*. **118,** 743–750.
19. Campisi, L., Yang, Y., Yi, Y., et al. (1999) Generation of enhancer trap lines in *Arabidopsis* and characterization of expression patterns in the inflorescence. *Plant J*. **17,** 699–707.

20. Galbiati, M., Moreno, M. A., Nadzan, G., Zourelidou, M., and Dellaporta, S. L. (2000) Large-scale T-DNA mutagenesis in *Arabidopsis* for functional genomic analysis. *Funct. Integr. Genomics* **1,** 25–34.
21. LeClere, S. and Bartel, B. (2001) A library of *Arabidopsis* 35S-cDNA lines for identifying novel mutants. *Plant Mol. Biol*. **46,** 695–703.
22. Goddijn, O. J., Lindsey, K., van der Lee, F. M., Klap, J. C., and Sijmons, P. C. (1993) Differential gene expression in nematode-induced feeding structures of transgenic plants harbouring promoter-gusA fusion constructs. *Plant J*. **4,** 863–873.
23. Koncz, C., Martini, N., Mayerhofer, R., et al. (1989) High-frequency T-DNA-mediated gene tagging in plants. *Proc. Natl. Acad. Sci. USA* **86,** 8467–8471.

12

Physical and Chemical Mutagenesis

Andrea Kodym and Rownak Afza

Summary

Important methods to artificially induce mutations are the use of chemical and physical agents. Most chemical mutagens are alkylating agents and azides. Physical mutagens include electromagnetic radiation, such as γ rays, X rays, and UV light, and particle radiation, such as fast and thermal neutrons, β and α particles.

Mutagenic treatment of seeds is the most convenient and, therefore, the standard method in seed propagated crops. Seeds can be treated in large quantities and are easily handled, stored, and shipped. It is fairly easy to repeat the conditions of mutagenic treatment, pre- and post-treatment, and hence, to obtain reproducible results within practical limits. Besides seed treatment, whole plants, cuttings, tubers, pollen, bulbs, corms, or in vitro plants or tissues can be treated.

This chapter is restricted to the commonly applied techniques of mutation induction in seeds by ethyl methanesulfonate (EMS) treatment and by γ and fast neutron irradiation.

Key Words

ethyl methanesulfonate, fast neutrons, γ radiation, mutation induction, mutagen sensitivity

1. Introduction

Mutagenesis is described as the exposure or treatment of biological material to a mutagen, i.e., a physical or chemical agent that raises the frequency of mutation above the spontaneous rate ***(1)***. Physical and chemical mutagens have been successfully used in plant breeding programs to artificially generate genetic variation for the development of new varieties with improved traits such as increased yield, earliness, reduced plant height, and resistance to disease ***(2)***. In recent years, mutation induction became also a powerful tool for the investigation of gene function and expression ***(3,4)***.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

Mutation is a random event at the single cell level. Hence, the population size of the M_1 and M_2 generation must be adequate to cope with the working objective. This size depends on the probability to generate the desired variation and on the inheritance of gene(s) ***(5)***. Some thousands of seeds are usually needed, being aware that the handling of such large populations requires efficient mass screening techniques.

Observations on M_1 plants show that with increasing dose, there is a reduction in germination or emergence, root length, seedling height, survival, and fertility ***(6)***. Delayed germination may be observed in mutagenized seeds as compared to control. When planting seeds in soil, emergence is taken as the criterion instead of germination. Germination is not a good indicator for an effective dose, because in the initial stage of germination mainly preformed organs are developing; a process that is fairly insensitive to mutagenesis. Only in the phase of active cell division the effects of mutation show clearly ***(7)***. Visible leaf spots are frequently generated following mutagen treatment in leguminous plants and also in other species.

Selection starts in the segregating M_2 population or in the M_3 for traits that can be screened for only on a row base. Dominant mutations, which are very rare, can be selected in M_1 already. (Nomenclature: M_0 seed refers to untreated seeds, M_1 seed to mutagenized seeds. M_1 plants are grown from M_1 seed, carrying M_2 seed. M_2 plants are grown from M_2 seed carrying M_3 seed, etc.)

1.1. Chemical Mutagens

Ethyl methanesulfonate (EMS; $CH_3SO_2OC_2H_5$) has been shown to be a very effective and efficient mutagen ***(8,9)*** and has probably become the most popular chemical mutagen ***(10)***. It is a colorless liquid compound with a molecular weight of 124 and is 8% soluble in water. EMS belongs to the group of the alkylating agents. These compounds have one or more reactive alkyl groups, which are capable of being transferred to other molecules at a position of higher electron density ***(11)***. According to their number of functional groups, they are mono-, bi-, or polyfunctional alkylating agents. Bi- and polyfunctional alkylating agents are generally more toxic than a monofunctional agent. EMS is a monofunctional alkylating agent.

Alkylating agents are very reactive, even with water. Hydrolysis (reaction with water) usually gives rise to compounds that are no longer mutagenic, but toxic to biological tissue. This means that the mutagen solution must be prepared just before use and never stored. The speed of hydrolysis is usually measured by the half-life or the time necessary for degradation to the half of the initial amount of alkylating agent. The half-life of EMS in water at pH 7.0 and at 20°C is 93 h, and at 30°C, the half-life is 26 h. EMS reacts with water as follows: $CH_3SO_2OC_2H_5 + H_2O \rightarrow CH_3SO_2OH + C_2H_5OH$

Since effects induced by alkylating agents are similar to those of ionizing irradiation, they are also classified as radiomimetic agents ***(12)***. Alkylation of DNA leads to the following reactions ***(13)***. Unstable triesters are formed, which release the alkyl group and interfere with DNA replication. Sometimes the phosphate triesters are hydrolyzed between sugar and phosphate, which result in the breakage of the DNA backbone. Alkylation of nitrogen bases occurs as well, as the reaction with guanine at the N-7 position is the most frequent event followed by adenine at N-3 and cytosine at N-1. Alkylated guanine is assumed to ionize differently than the normal guanine, and in such a way that guanine can pair with thiamine, thus leading to basepair errors. The alkalyted guanine can be separated from the deoxyribose leaving it depurinated. Depurination will leave a gap in the DNA template, thus, after replication, either a deletion will result, or any of the four bases may be inserted in the new strands opposite to the deletion.

One of the most crucial requirements for mutation induction is the selection of an efficient dose of the mutagenic agent for mutating the starting material. The dose can be defined as a particular mutagen concentration for a definite period of time at a particular temperature. If no relevant data are available, a preliminary experiment with different doses needs to be performed to determine the mutation effectiveness (mutations per unit dose) and mutation efficiency (ratio of mutation to injury or other effect). The mutagenic efficiency of a chemical mutagen depends not only on the properties of the chemical, but also on the genotype ***(14)***. Published data indicate that different species and even cultivars may respond differently. A quick and simple method to evaluate the mutagenic effect is to determine the primary injury in M_1 seedlings under greenhouse conditions. Primary injury includes reduction in seedling height, root length, survival, and fertility. It is advisable to perform a seedling test with a range of doses to determine the optimal treatment conditions for a specific cultivar and then to select a treatment in which growth reduction of about 20–30% was obtained ***(10)***.

An increase in concentration of EMS results in enhancing mutation, but causes proportionally even greater seedling damage or a decrease in survival. The vol of the treatment solution plays an important role. The vol needs to be large enough to avoid concentration gradients during treatment, in order to enable each seed to absorb the same number of moles of mutagen.

The duration of the treatment should be long enough to permit hydration and infusion of the mutagen to the target tissue. Experiments with labeled mutagens (EMS and methyl methanesulfonate [MMS]) showed that the uptake saturation in the embryo was dependent on the seed size, permeability of the seed coat, and contents of cell constituents ***(15)***. Mutagen saturation in the embryo may occur within 3–5 h in small seeds, while it may take up to 12 h in large

seeds if permeability does not limit the uptake of the mutagen. EMS produces strong acidic by-products upon hydrolysis, both inside the cell and in the mutagenic solution. If the treatment duration is too long, compared to the half-life of the mutagen, the EMS solution should be buffered or renewed with freshly prepared solution to reduce injury effects due to hydrolysis products and to maintain a constant mutagen concentration ***(12)***. The duration of the treatment can be shortened when using presoaked seeds.

Temperature does not directly affect the rate of diffusion, but it influences the rate of hydrolysis of the mutagenic solution. Since at low temperature the hydrolysis rate is decreased, the mutagen remains stable longer, thus ensuring reactivity with the target cells.

Modifying factors before, during, and after the treatment affecting the action of mutagens in a biological system include presoaking, hydrogen ion concentration, metallic ions, and storage conditions.

Presoaking of seeds enhances total uptake, the rate of uptake, and the distribution of the mutagen in the target tissue. The penetration of a maximum amount of mutagen into the embryo tissue, which is the actual target for mutagenic treatment, is enhanced. Wheat and barley embryo meristem tissue start DNA synthesis after 16–20 h of presoaking, which is one of the most sensitive stages that produces a high mutation frequency with relatively little chromosome damage ***(16)***.

Manipulation of the hydrogen ion concentration during and after the treatment plays an important role to obtain a favorable relationship between mutation yield and damage parameters. The ratio of gene mutation to chromosome mutation increases with the increase of pH of the EMS solution ***(17)***. The pH of the solution further effects hydrolysis of EMS. While the rate of hydrolysis of EMS seems not to be much affected by a low pH, the biological system is very sensitive at low pH. The pH of the solution should be monitored before and after the treatment ***(12)***.

The use of deionized water to prevent undesired effects by metallic ions is recommended. It is reported that certain metallic ions such as zinc and copper increase the frequency of chromosomal aberration induced by EMS ***(18)***.

Following EMS treatment, the storage conditions of the seeds can enhance injury. Mutation frequency and biological damage increase with increasing storage time, moisture content of the seed, applied mutagen dose, and temperature. An immediate dryback after the treatment leads to lethality, but this effect can be reduced or eliminated by post-treatment washing, whereby nonreacted chemicals and their hydrolytic by-products are rapidly removed. Prolonged post-treatment washing, followed by drying, leads to a reduction in biological damage without a decrease in mutation yield. Moreover, duration of the post-wash depends on the mutagenic dose applied, the temperature of the post-wash

solution, and the condition of the re-drying and storage. Barley seeds treated with a relatively high concentration of EMS, washed for 24 h at 24°C, re-dried, and stored at –20°C, showed no increase in damage ***(19)***. Storage of treated seeds with a moisture content of 12–14% is desirable for all purposes ***(20)***. Sowing should be implemented in wet soil to avoid artifacts (injury) due to extended dryback in dry soil.

1.2. Physical Mutagens

γ Rays are electromagnetic waves of very short wavelengths and are obtained by disintegration of the radioisotopes ^{60}Co or ^{137}Cs. γ sources can be installed in a γ cell, a γ room, or γ field. These are shielded by lead or concrete. Most γ sources are suitable for seed irradiation, as long as the size of irradiation space is sufficient and the dose rate allows practical irradiation times.

Fast neutrons are uncharged particles of high kinetic energy and are generated in nuclear reactors or in accelerators. The scientist should assess the feasibility for seed irradiation with the operators, since not all facilities are suitably equipped and can produce fast neutrons at a low degree of contamination with other radiation. As for example, a Standard Neutron Irradiation Facility (SNIF) has been especially constructed for swimming pool-type reactors to filter γ rays and thermal neutrons ***(21,22)*** and a Uranium-Shielded Irradiation Facility (USIF) can be used for TRIGA-type reactors ***(23)***. Casta ***(24)*** designed also a Standard Column Irradiation Facility (SCIF) that could be placed in thermal columns.

The two radiation types differ in their physical properties and, hence, in their mutagenic activity. γ Rays have a lower relative biological effectiveness (RBE) than fast neutrons, which implies that in order to obtain the same biological effect, a higher dose of γ radiation must be given ***(10)***. RBE is mainly a function of the linear energy transfer (LET), which is the transfer of energy along the ionizing track. γ Rays produce a few ionizations per micron of path (low LET) and belong to the category of sparsely ionizing radiation ***(12)***. Fast neutrons (high LET, densely ionizing radiation) impart some of their high kinetic energy via collisions, largely with protons within the material.

When radiation passes through tissue, physical events such as ionizations (ejection of electrons from molecules) and excitations (process of raising electrons to a higher energy state) occur and lead to effects in the DNA, membranes, lipids, enzymes, etc. Secondly, chemical events are induced that start with the formation of activated molecules, so-called free radicals ($OH^{\bullet}$ and $H^{\bullet}$) that arise from OH^- and H^+ ***(10)***. If oxygen is present, it reacts readily with radiation-induced free radicals to form peroxyradicals. In the case of low LET radiation, the formation of peroxyradicals is favored. In high LET radiation, the formation of hydrogen peroxide (H_2O_2) by recombination of free radicals

is favored. All radicals and hydrogen peroxide can react with biological molecules ***(25)***. Primary damage caused by radiation occurs randomly and is both physiological and genetic. Physiological recovery and repair of DNA are possible to some extent, as nondamaged molecules may take over metabolic processes and DNA repair mechanisms are activated.

Before starting any mutation induction studies, it is most crucial to select suitable doses. For mutation induction, it is advisable to use two to three doses along with a control ***(26,27)***. The applicable doses will depend on the breeding or research objective, the radiation type and the particular plant material. It is known that plant genera and species and, to a lesser extent, cultivars differ in their radiosensitivity ***(28)***. Radiosensitivity (radiation sensitivity) is a relative measure that gives an indication of the quantity of recognizable effects of the radiation exposure on the irradiated object ***(10)***. The radiosensitivity is influenced by biological factors (such as genetic differences, nuclear and interphase chromosome vol) and by environmental modifying factors (oxygen, water content, post-irradiation storage, and temperature) ***(29)***.

Modifying factors greatly affect mutagenic efficiency and reproducibility of results ***(12)***. Oxygen is the major modifying factor, while moisture content, temperature, and storage appear to be secondary, interacting with the oxygen effect. Oxygen shows a synergistic action with sparsely ionizing radiation, but oxygen effects during irradiation and post-irradiation storage can easily be prevented by adjustment of seed water content to 12–14%. The critical region is the embryo, but it can be assumed that the water content of the seed and the embryo of most species will be similar ***(29,30)***. Environmental factors are less important with densely ionizing radiation; thus, for fast neutron radiation, no seed moisture adjustment is necessary ***(29,31)***.

Unless data on the radiosensitivity of a given plant are already published (e.g., ***2,28***) or known from experience, the mutation induction program should be preceded by a radiosensitivity test. This is done by irradiating the seeds with a range of doses and by growing out the plants under greenhouse conditions. Radiosensitivity is assessed based on criteria such as reduced seedling height, fertility, and survival in the M_1 generation ***(6)***. A seedling height reduction of 30–40% is generally assumed to give a high mutation yield ***(6,14)***. The usefulness of radiation can be judged by mutagenic efficiency, which is the production of desirable changes free from association with undesirable changes ***(32,33)***. A high dose will increase mutation frequency (the frequency at which a specific kind of mutation or mutant is found in a population of cells or individuals ***[34]***), but will be accompanied by negative features, such as sterility. When selecting the doses, it will be necessary to find a treatment regime providing high mutagenic efficiency.

Precise dosimetry data should be available for the γ sources, in order to calculate the time of exposure needed to obtain a given dose and to define the homogeneity of the radiation field. Dosimetry of γ rays is comparatively simple and can be done, e.g., using Fricke dosimetry ***(35)***, in which the oxidation of Fe^{2+} to Fe^{3+} as a result of radiation is measured spectrophotometrically. Alternatively, information on the dose rate can be found in the manual of the γ source. Manufacturers of γ cells further supply isodose curves for the sample chamber, because the radiation field is usually heterogeneous. For γ rooms and γ fields, the inverse-square-distance law is relevant: intensity decreases proportional to the inverse square of the distance ($1/r^2$) from a point source (double distance results in quarter intensity). For fast neutron radiation, dosimetric measurements have to be done during each radiation treatment, e.g., by performing the sulfur threshold detector method ***(36)***, since the neutron flux in the seed irradiation unit is not constant.

The Gray (symbol Gy), the SI (Système Internationale) unit used to quantify the absorbed dose of radiation (1 Gy = 1 J/kg) replaced the old unit rad; 1 Gy = 100 rads or 1 krad = 10 Gy. The absorbed dose rate (Gy/s or Gy/min) indicates how much energy the irradiated material absorbs during a given unit of time. The length of exposure and the dose rate determines the radiation dose. Exposure during short times (s to a few h) at a high dose rate is referred to as acute and is most applied in irradiation programs. Exposure for a prolonged period of time (d to mo) at a low dose rate is called chronic ***(10,12)***. γ Cells are commonly used for acute irradiation, γ rooms and γ fields are used for chronic irradiation. An alternative to chronic irradiation is the split dose irradiation in which more than one irradiation is carried out, interrupted by (one or more) time intervals. In general it appears that the biological damage after fractionation of a total dose is less than after application of a single acute dose, due to recovery processes ***(10)***.

2. Materials

2.1. EMS Treatment

1. Seeds of barley (*Hordeum vulgare* L.).
2. EMS ($C_3H_8O_3S$) (Sigma, cat. no. M0880).
   Note: Mutagenic. Store the original EMS always in an airtight colored bottle, preferably inside a sealed chamber containing a desiccant. Make solution fresh as required.
3. Mesh bags (polyethylene) (size approx 11 × 7 cm), which can be made from plastic screen available in the market.
4. Deionized or distilled water.
5. Fume hood.

2.2. γ Radiation

1. Seeds of barley (*Hordeum vulgare* L.).
2. Paper or mesh bags (polyethylene), which can be made from plastic screen available in the market.
3. Vacuum desiccator with 60% glycerol: distilled water mixture (v/v).
4. γ source. An experienced operator should handle the facility. Depending on the γ source, it may be of high or low radiation hazard, and a dosimeter (e.g., thermoluminescence dosimeters [TLD]) to monitor exposure to radiation may need to be worn.

2.3. Fast Neutron Radiation

1. Seeds of barley (*Hordeum vulgare* L.).
2. Nuclear reactor or accelerator suited for fast neutron seed irradiation. Radiation hazard, experienced staff will perform the irradiation.

3. Methods (*see* Notes 1 and 2)

3.1. EMS Treatment

1. Take dry, quiescent barley seeds with high germiability. Seeds should be of high quality and genetically as uniform as possible to avoid impurity in starting material. Remove any injured, diseased, or atypical seeds (*see* **Note 3**).
2. Choose three doses of EMS in the range of 0.05–0.1 M solution, temperature 30°–35°C and duration 0.5–2 h ***(37)***. Include a small amount of seeds as control treatment (not to be mutagenized) (*see* **Notes 4–6**).
3. Place seeds in mesh bags. The number of bags depends on the number of treatments. Fold the tops of the bags over and close with a plastic paper clip; attach a cotton string and tag with treatment identification.
4. Place the bags in a beaker with distilled (or deionized) water and soak the seeds for 16–20 h at 20°–22°C. During presoaking, intermittent shaking or bubbling with air or oxygen needs to be done to provide good aeration (*see* **Note 7**).
5. Take the seed out of the water and shake off excess water.
6. Just before use, prepare the EMS solution of the desired concentration using distilled or deionized water in a fume hood. To prepare 100 mL of a 0.1 *M* solution use 1.0615 mL of commercially available EMS solution (d = 1.17 g/mL). EMS shall be vigorously shaken, e.g., in a bottle to achieve a homogenous emulsion. Use at least 1 mL of solution per seed (*see* **Notes 8–16**).
7. The seeds are then subjected to mutagenic treatment of the desired concentration and duration using a water bath (*see* **Notes 17** and **18**).
8. To wash, place the bags of seeds into a bucket under running cold tap water for a few hours to remove traces of EMS in the seed embryo.
9. Dispose of the unused EMS mutagen solution by adding 4% NaOH or 10% sodium thiosulfate in large excess. Pour into a container, which is marked with "disposal of suspected carcinogen" and let stand for at least six half-lives. Half-

life of EMS in 4% NaOH is 6 h at 20°C and 3 h at 25°C. For EMS in a 10% sodium thiosulfate solution, the half-life is 1.4 h at 20°C and 1 h at 25°C.

10. Take the bags of seeds, shake off excess moisture, and place the seeds onto blotting paper for a short while to surface-dry. It is best to sow seeds immediately after treatment, to minimize artifacts, in a well-prepared seedbed. This is called "wet treatment." A dry soil should be irrigated after sowing of mutagenized seeds to avoid injury by dryback in soil. The environmental conditions, as well as water and nutrient supply should be uniform for all treatments (*see* **Note 19**).
11. For storing or transporting seeds, seeds should be dry. To dry, hang the bags of seeds in an air current (called dryback treatment). After 1 to 2 d of drying, store the seeds in a refrigerator.

3.2. γ Radiation

1. Take dry quiescent barley seeds with high germiability. Seeds should be of high quality and genetically as uniform as possible to avoid impurity in starting material. Remove any injured, diseased, or atypical seeds (*see* **Note 3**).
2. Select two different doses in the range of 100–250 Gy and include also a small amount of seeds as control treatment (not to be irradiated) (*see* **Notes 4–6**).
3. Pack seeds in lots according to the doses in mesh bags or water permeable paper bags and label them with species and variety name, date and dose. The size of the bags must not exceed the size of the irradiation facility (*see* **Note 20**).
4. Place seeds in a vacuum desiccator over 60% glycerol-distilled water mixture at room temperature for a minimum of 7 d. In the desiccator, the relative humidity should be about 73%, which can be checked using a hygrometer (*see* **Notes 21–23**).
5. The moisture-equilibrated seeds are irradiated using a γ source. Only persons familiar with operation and radiation safety should use the equipment. γ Irradiation is generally performed by an experienced operator of the γ source who knows its characteristics and calculates the present dose rate and required exposure time.
   a. Calculate the present dose rate (D_t) from the dose rate (D_o) provided by the supplier in the manual or from dosimetry data (based on refs. ***38*** and ***39***):

      $D_t = D_o \times e^{-\lambda \times t}$ — D_t = dose rate of today, present dose rate; D_o = dose rate at time 0, initial dose rate; λ = decay constant; t = elapsed time since determination of D_o

      $\lambda = \ln 2/T$ — T = physical half-life (^{60}Co = 5.27 a = 1924 d, ^{137}Cs = 30 a)

      $D_t = D_o \times e^{-\ln 2 \times t/T}$

      $D_t = D_o \times 2^{-t/T}$

   b. Calculate the required exposure time (min, s) based on the present dose rate (Gy/s or Gy/min) of the γ source.

      Exposure time = desired dose/dose rate (D_t)

c. The seeds should be irradiated in a homogenous field. Check the manual of the γ cell for isodose line records and accordingly raise the seeds into the homogenous position by using for example Styrofoam disks (*see* **Note 24**).

6. γ Irradiation does not leave any radioactivity in the treated seeds ***(10)***, and they can be handled without precautions.
7. Plant seeds in a well-prepared seedbed, meeting the requirements of the plant crop. It is advised to sow the seeds as soon as possible after radiation treatment ***(10)***. The environmental conditions, as well as water and nutrient supply, should be uniform for all treatments (*see* **Note 19**).

   If necessary, store the seeds dry for 2–4 wk at room temperature ***(40)***. For extended periods: store seeds dry, as much as possible in the absence of oxygen, i.e., sealed in airtight bags or vials, or in the dark or at 2°C or –5°C ***(10,12)*** to slow down metabolic activity.

3.3. Fast Neutron Treatment

1. Take dry quiescent barley seeds with high germiability. Seeds should be of high quality and genetically as uniform as possible to avoid impurity in starting material. Remove any injured, diseased, or atypical seeds (*see* **Note 3**).
2. Select two different doses in the range of 3–6 Gy and also include a small amount of seed as control treatment (not to be irradiated) (*see* **Notes 4–6**).
3. Pack the seeds to be irradiated in lots according to the dose in airtight plastic bags or vials (to prevent potential spoilage with pool water) and label them with species, variety name, and the dose. The bags or vials must not exceed the size of the irradiation facility.
4. Hand the seeds over to the staff operating the nuclear reactor or accelerator who will then perform the irradiation and dosimetric measurements. The absorbed dose should lie within ±5% of the desired dose.
5. The containers are left after irradiation for a few days, to "cool off," since fast neutron radiation normally causes a low level of temporary activity ***(10)***.
6. Plant seeds in a well-prepared seedbed, meeting the requirements of the plant crop. It is advised to sow the seeds as soon as possible after radiation treatment ***(10)***. The environmental conditions, as well as water and nutrient supply, should be uniform for all treatments (*see* **Note 19**). If necessary, store the seeds dry for 2–4 wk at room temperature ***(40)***. For extended periods: store seeds dry, as much as possible in the absence of oxygen, i.e., sealed in airtight bags or vials, or in the dark or at 2°C or –5°C ***(10,12)*** to slow down metabolic activity.

4. Notes

1. There is no single ideal protocol for seed treatment of different species and even different varieties. A resume of conditions considered most effective for inducing mutations given here shall serve as guidelines. Each genotype must be tested for the optimal treatment conditions within its range of conditions.

Table 1
Successful Doses of γ Rays and Fast Neutrons

Plant name	γ rays	Fast neutrons
Arabidopsis[a]		60 Gy
Brassica napus[a]	600–800 Gy	
Glycine max	100–200 Gy	
Triticum aestivum	100–300 Gy	3–6 Gy
Oryza sativa	200–400 Gy	10–20 Gy
Sorghum sp.	250 Gy	
Zea mays	250 Gy	

[a]*Brassicaceae* are generally fairly radiation-tolerant.

2. Detailed records should be kept to be able to repeat a successful experiment.
3. In dormant seeds, it is necessary to overcome dormancy mechanisms before undergoing mutation induction by, e.g., cold treatment (prechilling), a period of heat, or breaking of a hard seed coat (scarification).
4. Decide on two to three doses based on published data, experience, or prior mutagen sensitivity test results. Successful doses for various mutagens that lead to the selection of a mutant are found on (http://www-mvd.iaea.org) ***(2)***. **Table 1** is referring to some successful doses of γ rays and fast neutrons ***(3,26,41)*** that can serve as guidelines.
5. Use populations of sufficient size to assure success (formulas for calculating the population size are discussed in ***[5,10,13]***). In general, aim at an M_2 population of 5000–10,000.
6. A nonmutated control should always be included. It is treated like the material to be mutagenized except for the mutagen exposure. The control population should be grown to provide for comparisons and to assess the phenotypic variation of the parent stock ***(42)***.
7. Presoak seeds to bring the embryos to G1 phase or the onset of S phase. The time period for presoaking depends on the seed coat. Seeds having a hard and thick seed coat require longer time than others with a thin and soft seed coat.
8. EMS should be labeled clearly with "carcinogen" or "biohazard" and should always be kept in a closed container. Any access of water will destroy it via hydrolysis.
9. All the handling of the EMS, from the opening of the bottle, preparation of solutions, treatment of biological material, washing of glassware, etc., must be confined in a limited area covered with filter paper, if possible in a hood with good ventilation. When no hood is available, use a face mask. Absorbing materials such as filter paper or sawdust can be used for cleaning accidentally contaminated areas ***(43)***.
10. Eating, drinking, and smoking in the mutagen laboratory are strictly prohibited.
11. To avoid skin contact, use disposable hand gloves. Dispose of the gloves when

you see any sign of contamination. All persons shall wash their hands immediately after completion of experiments.

12. Protective clothing (laboratory coat) must be worn. Use pipeting aids and, if available, disposable pipets. Precaution needs to be taken for the avoidance of potential contamination of pipeting aids. Collect gloves, disposable pipets, etc., separately as toxic waste.
13. Contaminated glassware and glass pipets should be put into 10% sodium thiosulfate ($Na_2S_2O_3 \cdot 5H_2O$) solution for inactivation. The materials should stay in the thiosulfate solution for at least six half-lives.
14. If skin gets contaminated, wash with detergent and a large quantity of water and neutralize with 10% sodium thiosulfate.
15. Clothing contaminated with EMS shall first be decontaminated with 10% sodium thiosulfate solution and excess water.
16. A minimum of 0.5–1.0 mL/seed is used for the majority of cereals, while large-seeded grain legumes require at least 2 mL/seed.
17. Mutagenic treatment can be performed at room temperature or, by making use of a water bath, at a higher temperature.
18. Where the treatment duration is too long compared to the half-life of the mutagen, the EMS solution should be buffered or renewed with freshly prepared solution when approx one-fourth of the mutagen has been hydrolyzed. The half-life of EMS in water at pH 7.0 and 20°C is 93 h and at 30°C is 26 h. If a buffer is used in the solution, phosphate buffer at strength of maximum 0.1 *M* at pH 7.0 is recommended to avoid injury effects on treated seeds.
19. Field experiments are more difficult to maintain, and survival is lower than under greenhouse conditions, because of biotic and abiotic stress and varying environmental conditions. One may be confronted by problems with volunteer crop plants and bird, insect, and rat damage. Therefore, under field conditions, crop, pest and disease, and water management must be optimized.
20. For the seed moisture adjustment prior to γ radiation, the seed coat must be water-permeable. Otherwise, it must be removed or mechanically or chemically modified. Seed coats can be rubbed with sandpaper, nicked with a knife, or filed with a metal file.
21. Up to 500 g of seeds can be adjusted in one desiccator if 1000 mL of glycerol/water mixture are used ***(29,30)***. Greater quantities and/or larger seeds (larger than cereals) are adjusted for a minimum of 14 d.
22. Different species may not equilibrate to the same water content at a particular relative humidity ***(44)***. In general, it is sufficient to follow the desiccation procedure described above and to have a standardized method without determining the seed water content. However, if uncertainties arise, the seed water content can be measured according to the rules of the International Seed Testing Association, Official Grain Standards of the United States, or official methods of the Association of Official Agricultural Chemists ***(33,45)***.
23. If γ radiation does not immediately follow the moisture equilibration, the seeds should be packed in airtight plastic bags to maintain the desired moisture content of 12–14%.
24. Irradiating seeds in the homogenous field of the γ source is giving the most accu-

rate dose. However, when treating large quantities of seeds in small radiation facilities, using only the homogenous radiation space may not be feasible.

Acknowledgments

The authors would like to thank Dr. Helmut Brunner for sharing his vast experience in mutation induction and the critical reading of the manuscript.

This work is based on protocols established at the Plant Breeding Unit, Agency's Laboratories Seibersdorf, International Atomic Energy, Austria; the authors wish to acknowledge the input of all staff involved past and present.

References

1. Rieger, R., Michaelis, A., and Green, M. (1976) *Glossary of Genetics and Cytogenetics*. Springer Verlag, New York.
2. Maluszynksi, M. (2001) Officially released mutant varieties—The FAO/IAEA Database. *Plant Cell Tissue Organ Cult.* **65,** 175–177.
3. Li, X., Song, Y., Century, K., et al. (2001) A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.* **27,** 235–242.
4. Mittelsten Scheid, O., Afsar, K., and Paszkowski, J. (1998) Release of epigenetic gene silencing by trans-acting mutations in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **95,** 632–637.
5. Brock, R. D. (1977) When to use mutations in plant breeding, in *Manual on Mutation Breeding*. IAEA Technical Reports Series, No. 119, 2nd ed., pp. 213–219.
6. Brunner, H. (1995) Radiation induced mutations for plant selection. *Appl. Radiat. Isot.* **46,** 589–594.
7. Micke, A. (1997) Mutagenese und Induzierte Mutationen in der Züchtung, in *Biologische Grundlagen der Pflanzenzüchtung* (Odenbach, W., ed.), Parey Buchverlag, Berlin, pp. 218–239.
8. Heslot, H., Ferrary, R., Levy, R., and Monard, C. (1959) Recherches sur les substances mutagenes (halogeno 2–ethyle)amines, derives oxygenes du sulfure de bis (chloro-2–ethyle), ester sulfoniques et sulfuriques, C.R. Seanc. Hebd. Acad. Sci., Paris 248, 729.
9. Heslot, H., Ferrary, R., Levy, R., and Monard, C. (1961) Induction de mutations chez l'orge. Efficacite relatives des rayons gamma, du sulfate d'ethyle, du methane sulfonate d'ethyle et de quelques autres substances. Effects of ionizing radiation on seeds (Proc. Conf. Karlsruhe, 1960), IAEA, Vienna, 243–250.
10. Van Harten, A. M. (1998) *Mutation Breeding Theory and Practical Applications*. Cambridge University Press, Cambridge.
11. Heslot, H. (1977) Review of main mutagenic compound, in *Manual on Mutation Breeding*. IAEA Technical Reports Series, No. 119, 2nd ed., pp. 51–58.
12. Brunner, H. (1991) Methods of induction of mutations, in *Advances in Plant Breeding, Vol. 1.* (Mandal, A. K., Ganguli, P. K., and Banerjee, S. P., eds.), CBS Publishers, Delhi, pp. 187–252.
13. Siddiqui, B. A. and Khan, S. (1999) Mutagenesis: tools and techniques—a practi-

cal view, in *Breeding in Crop Plants: Mutations and In Vitro Mutation Breeding*. Kalyani Publishers, Ludhiana, New Delhi, pp. 20–32.
14. Kamra, O. P. and Brunner, H. (1977) Chemical mutagens: dose, in *Manual on Mutation Breeding*, IAEA Technical Reports Series, No. 119, 2nd ed., pp. 66–69.
15. Brunner, H. and Ashri, A. (1986) Dynamics of mutagen uptake of EMS and MMS mutagenised sees of peanut and sesame. Proc. Int. Symp. On new genetical approaches to crop improvement. Karadu, Pakistan, pp. 217–227.
16. Mikaelsen, K., Ahnstroem, G., and Li, W. C. (1968) Genetic effects of alkylation agents in barley. Influence of post treatment-storage, metabolic state, and pH of the mutagenic solution. *Heriditas* **59,** 353–374.
17. Ramanna, M. S. and Natarajan, A. T. (1965) Studies on the relative mutagenic efficiency of alkylating agents under different conditions of treatment. *Indian J. Genet. Plant Breed* **25,** 24–25.
18. Moutschen-Dahmen, J. and Moutschen-Dahmen, M. (1963) Influence of Cu^{2+} and Zn ions on the effects of ethylmethane sulfonate EMS on chromosomes. *Experientia* **19,** 144–147.
19. Sato, M. and Gaul, H. (1967) Effect of ethylmethane sulfonate on the fertility of barley. *Rad. Bot.* **7,** 7.
20. Gaul, H., Frimmel, G., Gichner, T., and Ulonska, E. (1972) Efficiency of mutagenesis, in *Induced Mutations and Plant Improvement* (Proc. Meeting Buenos Aires, 1970), IAEA, Vienna, pp. 121–139.
21. Burtscher, A. (1968) Experience with the standard neutron irradiation facility in the ASTRA reactor, in *Neutron Irradiation of Seeds II*. Technical Reports Series No. 92, IAEA, Vienna, pp. 97–106.
22. Burtscher, A. and Casta, J. (1967) Facility for seed irradiations with fast neutrons in swimming-pool reactors: a design study, in *Neutron Irradiation of Seeds*. Technical Reports Series No. 76, IAEA, Vienna, pp. 41–61.
23. Casta, J. (1968) Facility for seed irradiation with fast neutrons in TRIGA type reactors, in *Neutron Irradiation of Seeds II*. Technical Reports Series No. 92, IAEA, Vienna, pp. 113–121.
24. Casta, J. (1972) Facility for fast neutron irradiation in thermal columns, in *Neutron Irradiation of Seeds III*. Technical Reports Series No. 141, IAEA, Vienna, pp. 105–112.
25. Ahnström, G. (1977) Radiobiology, in *Manual on Mutation Breeding*. IAEA Technical Reports Series, No. 119, 2nd ed., pp. 21–27.
26. Kamra, O. P. (1997) Internal report, Special Service Agreement. IAEA, Vienna.
27. International Atomic Energy Agency, Agricultural Biotechnology Laboratory, Seibersdorf (1985) Mutation induction in plants by ionizing radiation, Video film, Vienna.
28. Brunner, H. (1977) Radiosensitivity of a number of crop species to gamma and fast neutron radiation, in *Manual on Mutation Breeding*. IAEA Techn. Reports Series, No. 119, 2nd ed., pp. 44–45.
29. Conger, B. V., Konzak, C. F., and Nilan, R. A. (1977) Radiation sensitivity and modifying factors and methods of applying pre- and post-treatments, in *Manual on Mutation Breeding*. IAEA Techn. Reports Series, No. 119, 2nd ed., pp. 40–50.
30. Konzak, C. F., Bottino, P. J., Nilan, R. A., and Conger, B. V. (1968) Irradiation

of seeds, a review of procedures employed at Washington State University, in *Neutron Irradiation of Seeds II*. Technical Reports Series, No. 92, IAEA, Vienna, pp. 83–95.
31. Mikaelsen, K. and Kramer, J. (1972) Effects of water content, oxygen and metabolic state on genetic effect of fast neutrons and gamma radiation of barley seeds, in *Neutron Irradiation of Seeds III*. Technical Reports Series No. 141, IAEA, Vienna, p. 59.
32. Konzak, C. F., Nilan, R. A., Wagner J., and Foster, R. J. (1965) Efficient chemical mutagenesis. *Suppl. Rad. Bot.* **5,** 51.
33. Nilan, R. A., Konzak, C. F., Wagner, J., and Legault, R. R. (1965) Effectiveness and efficiency of radiations for inducing genetic and cytogenetic changes. *Suppl. Rad. Bot.* **5,** 71–89.
34. Suzuki, D. T., Griffiths, A. J. F., Miller, J. H., and Lewontin, R. C. (1989) *An Introduction to Genetic Analysis*, 4th ed. W.H. Freeman and Company, New York.
35. Fricke, H. and Hart, E. J. (1966) Chemical dosimetry, in *Radiation Dosimetry, 2nd ed., Vol. 2*. Academic Press, New York, pp. 167–239.
36. International Atomic Energy Agency (1968) Recommendations. Annex II, in *Neutron Irradiation of Seeds II*. Technical Reports Series, No. 92, IAEA, Vienna, pp. 150–162.
37. Mikaelsen, K. (1977) Ethyl methanesulfonate treatment of cereal seeds (barley), in *Manual on Mutation Breeding*. IAEA Technical Reports Series, No. 119, 2nd ed., pp. 78–79.
38. Brunner, H. (1997) Calculation of dose rate from a Co-60 source. Interregional training course handout, FAO/IAEA Laboratories, Seibersdorf, Austria.
39. Feldmann, U. (1989) Atomic physics and radioactivity: brief outline. Handout, Entomology Unit, FAO/IAEA Laboratories, Seibersdorf, Austria.
40. Conger, B. V., Nilan, R. A., Konzak, C. F., and Metter, S. (1966) The influence of seed water content on the oxygen effect in irradiated barley seeds. *Radiat. Bot.* **6,** 129–144.
41. Bathia, C. R., Nichterlein, K., and Maluszynski, M. (1999) Oilseed cultivars developed from induced mutations and mutations altering fatty acid composition. *Mutation Breeding* Review 11.
42. Konzak, C. F. and Mikaelsen, K. (1977) Selecting parents and handling the M_1–M_3 generations for the selection of mutants, in *Manual on Mutation Breeding*. IAEA Technical Reports Series, No. 119, 2nd ed., pp. 125–138.
43. Brunner, H. (1985) Standards for laboratory operations involving chemical mutagens. A training manual, Plant Breeding Unit, FAO/IAEA Laboratories, Seibersdorf, Austria.
44. Osborne, T. S. and Lunden, A. O. (1965) Prediction of seed radiosensitivity from embryo structure, in *The Use of Induced Mutations in Plant Breeding* (Rep. FAO/IAEA Techn. Meeting Rome, 1964), Pergamon Press, Oxford, pp. 133–149.
45. Zeleny, L. (1961) Ways to test seeds for moisture, in *Seeds*, in The Yearbook of Agriculture, USDA, Washington, pp. 443–447.

13

High-Throughput TILLING for Functional Genomics

Bradley J. Till, Trenton Colbert, Rachel Tompa, Linda C. Enns, Christine A. Codomo, Jessica E. Johnson, Steven H. Reynolds, Jorja G. Henikoff, Elizabeth A. Greene, Michael N. Steine, Luca Comai, and Steven Henikoff

Summary

Targeting-induced local lesions in genomes (TILLING) is a general strategy for identifying induced point mutations that can be applied to almost any organism. Here, we describe the basic methodology for high-throughput TILLING. Gene segments are amplified using fluorescently tagged primers, and products are denatured and reannealed to form heteroduplexes between the mutated sequence and its wild-type counterpart. These heteroduplexes are substrates for cleavage by the endonuclease CEL I. Following cleavage, products are analyzed on denaturing polyacrylamide gels using the LI-COR DNA analyzer system. High-throughput TILLING has been adopted by the *Arabidopsis* TILLING Project (ATP) to provide allelic series of point mutations for the general *Arabidopsis* community.

Key Words

TILLING, mutation, CEL I, reverse genetics, functional genomics

1. Introduction

Targeting-induced local lesions in genomes (TILLING) is a reverse genetic method that combines random chemical mutagenesis with polymerase chain reaction (PCR)-based screening of gene regions of interest ***(1,2)***. This provides a range of allele types, including missense and knock-out mutations, which are potentially useful in a variety of gene function and interaction studies. TILLING is especially suitable for plants, even for those that lack well-developed genetic tools. We have developed a TILLING protocol that achieves high-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

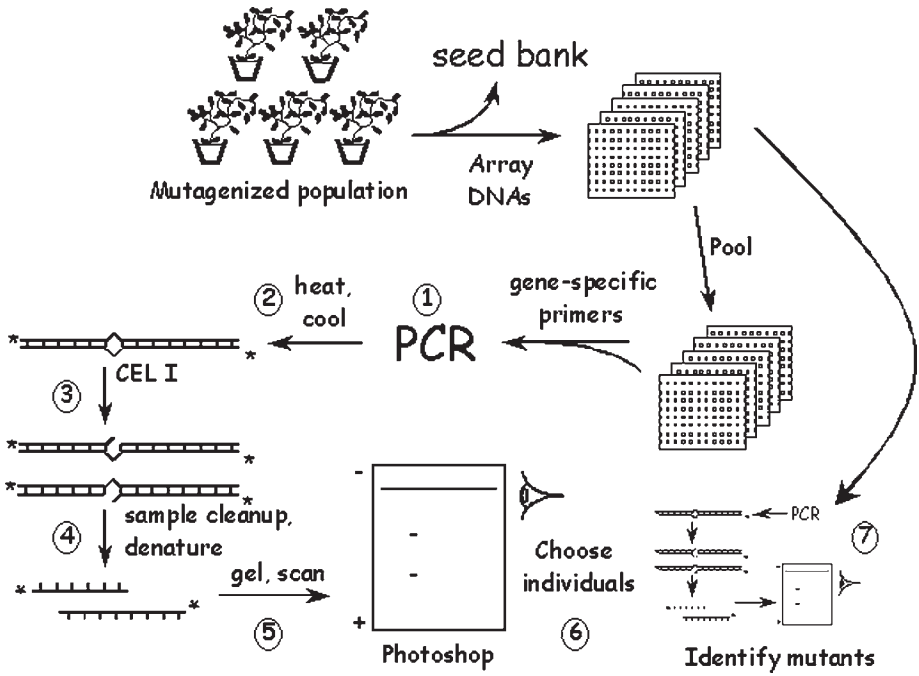


Fig. 1. Outline of high-throughput TILLING. DNA from individuals from a mutagenized population are first arrayed in a 96-well format. Seeds from these individuals are collected and stored for later analysis. DNAs are then pooled up to eight-fold to increase screening throughput. High-throughput TILLING can be separated into seven steps as noted here: (1) PCR is performed on pooled populations with target-specific primers labeled with fluorescent dyes; (2) an extended incubation at 99°C both kills *Taq* and denatures PCR products, followed by a slow cooling step, in which PCR products reanneal forming heteroduplexes; (3) heteroduplexes are digested with the nuclease CEL I; (4) samples are passed through a Sephadex G50 spin plates to remove salts and buffer components that are inhibitory to gel runs and laser detection; (5) samples are loaded onto 100-tooth membrane combs, and samples are electrophoresed; (6) gel images are analyzed for mutations in pools; and (7) mutations are tracked down to the individual.

throughput using gel-based screening of heteroduplex PCR products that have been preferentially cleaved at mismatches ***(3)***.

The general scheme for high-throughput TILLING is outlined in **Fig. 1**. DNAs from mutagenized individuals are first arrayed in a 96-well format. Samples are then pooled to increase throughput. The initial screening procedure consists of: *(i)* setting up and running the PCR on pooled DNAs using IRD700 and IRD800 primers for IR^2 gel analysis (LI-COR) ***(4)***; *(ii)* heat inactivation of polymerase and annealing to create hetereoduplexes; *(iii)* CEL I

digestion of heteroduplexes; *(iv)* sample cleanup on G50 spin plates; *(v)* loading and running the gels; *(vi)* processing and examining the gel images to identify mutations; and *(vii)* repeating steps *i–vi* on individuals identified in the pooled screen. This is followed by sequencing the mutant region to ascertain the mutation. The screening strategy of labeling at both ends provides confirmation of each band detected, as its complement is detected independently and allows for screening of 1-kb fragments **(Fig. 2)**. In addition, running two channels on the same gel simplifies comparisons and helps to identify artifactual primer–dimer bands, which appear in both channels.

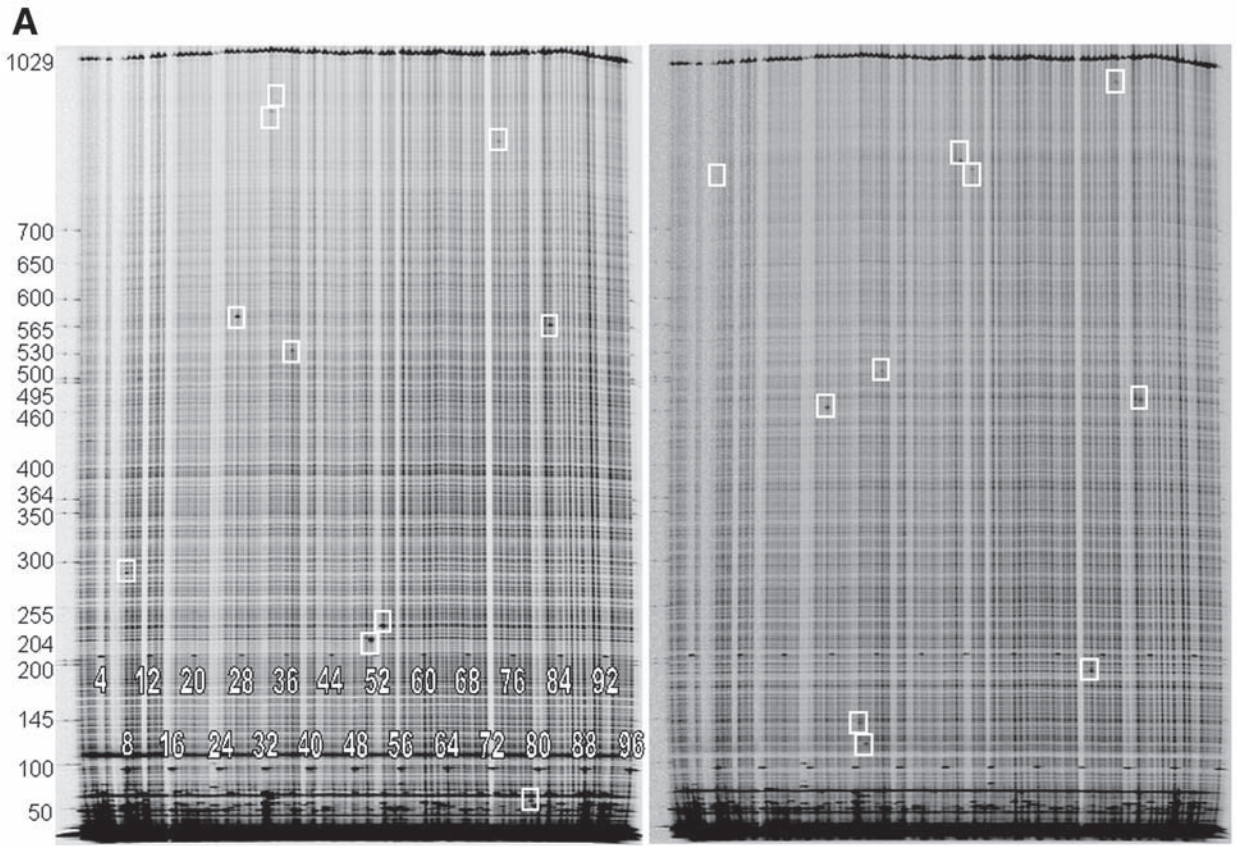
2. Materials

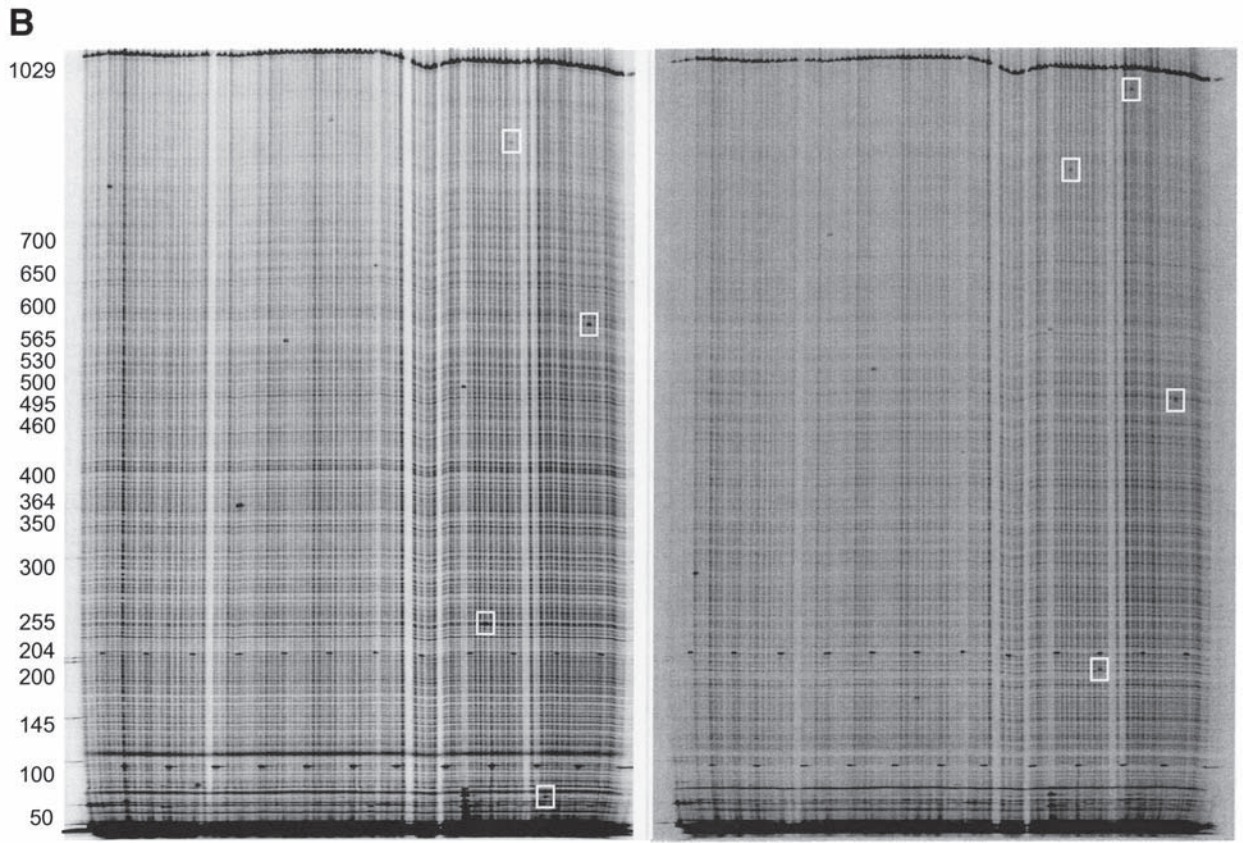
2.1. PCR

1. Ex-*Taq* DNA polymerase (Takara). Store at –20°C.
2. 10× Ex-*Taq* PCR buffer (Takara) supplied with Ex-*Taq*. Store at –20°C.
3. 2.5 m*M* (each) dNTPs (Takara) supplied with Ex-*Taq*. Store at –20°C.
4. 25 m*M* $MgCl_2$.
5. TE: 10 m*M* Tris-HCl, 1 m*M* ethylene diamine tetraacetic acid (EDTA), pH 7.4.
6. Left primer (melting temperature [T_m] 70°C) labeled 5' with IRD700 (MWG) 100 μ*M* in TE. Store at –80°C.
7. Left primer (T_m 70°C) unlabeled (MWG) 100 μ*M* in TE. Store at –80°C.
8. Right primer (T_m 70°C) labeled 5' with IRD800 (MWG) 100 μ*M* in TE. Store at –80°C.
9. Right primer (T_m 70°C) unlabeled (MWG) 100 μ*M* in TE. Store at –80°C.
10. Primer mixture: 3 μL IRD700 left primer, 2 μL unlabeled right primer, 4 μL IRD800 right primer, 1 μL unlabeled right primer. Store at 4°C and discard after 1 wk (*see* **Note 1**).
11. PCR mixture for a 96-well plate: 360 μL water, 57 μL 10× Ex-*Taq* buffer, 68 μL 25 m*M* $MgCl_2$, 92 μL 2.5 mM dNTP mixture, 4 μL primer mixture, 6 μL Ex-*Taq*. Mix on ice, adding polymerase last. Use immediately and discard remainder after use.

2.2. CEL I Digestion

1. 10× CEL I buffer: 5 mL 1 *M* $MgSO_4$, 5 mL 1 *M* 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH 7.5, 2.5 mL 2 *M* KCl, 100 μL 10% Triton® X-100, 5 μL 20 mg/mL bovine serum albumin, 37.5 mL water. Store aliquots at –20°C.
2. Cel I reaction mixture for a 96-well plate: 2.4 mL water, 420 μL 10× CEL I buffer, 36 μL CEL I. The amount of CEL I may vary based on the prep. Mix on ice, use immediately, and discard remainder after use.
3. Stop solution: 0.15 *M* EDTA, pH 8.0.


A
1029
700
650
600
565
530
500
495
460
400
364
350
300
255
204
200
145
100
50
4 12 20 28 36 44 52 60 68 76 84 92
8 16 24 32 40 48 56 64 72 80 88 96


B
1029
700
650
600
565
530
500
495
460
400
364
350
300
255
204
200
145
100
50

2.3. Spin Plate Cleanup

1. 96-Well membrane plates (Millipore).
2. Sephadex® G50 medium (Amersham Pharmacia Biotech).
3. Sephadex spin plates: use a Sephadex loading device (MultiScreen 45-μL column loader; Millipore; cat. no. MACL 096 45) to fill all wells of a 96-well membrane plate with approx 0.03 g of G50. Hydrate with 300 μL water. Allow the plate to stand for 1 h at room temperature. Plates may be stored at 4°C in a sealed container to prevent evaporation for up to 1 wk.
4. Deionized formamide: add 12.5 g of deionizing resin (AG® 501-X8 Resin; Bio-Rad) to 250 mL formamide. Stir for 1 h and filter through Whatman no. 4 filter paper (Whatman).
5. Formamide load buffer: 250 mL deionized formamide, 5 mL 0.5 *M* EDTA, pH 8.0, 60 mg bromphenol blue. Store at room temperature.
6. Sample receptacle plate: 96-well PCR plate containing 1.5 μL formamide load buffer per well. Plates may be made up to 4 wk in advance and stored at room temperature.
7. 95- and 200-bp markers: perform PCR (**Subheading 3.1.**) and spin plate cleanup (**Subheading 3.3.**) with primers designed to yield a 95- or 200-bp fragment.

2.4. Electrophoresis

1. Ammonium persulfate (APS): dissolve APS 10% (w/v) in water. Store at –20°C in small aliquots avoiding repeated freeze–thaw cycles.
2. Combine 20 mL gel mixture (6.5% acrylamide, 7 *M* urea; LI-COR), 15 μL N,N,N',N'-tetramethyl ethylenediamine (TEMED), and 150 μL APS. Use immediately.
3. 0.8× TBE buffer: dissolve 89.2 g Tris-base and 45.8 g boric acid in water, add 68 mL 0.25 *M* EDTA, pH 8.0, bring vol to 10 L.
4. IR^2 gel analyzer, 25-cm glass plates, and 25-mm spacers (LI-COR).
5. 100-tooth membrane combs (The Gel Company).
6. Size Standard IRDye™ 700 and IRDye 800 molecular weight markers (LI-COR).

Fig 2. Complementary strand images of an eight-fold pool screening gel **(A)** and an unpooled individual gel **(B)**. **(A)** Nine mutations (white boxes) were identified in this pool screening gel. The IRD700 (left) and IRD800 (right) channels are shown. All nine mutations were confirmed by the appearance of bands in both channels that add up to the full-length product of 1029 bp. Lanes were identified using the 95- and 200-bp markers present in every 4th lane, as indicated on the IRD700 image in panel A. Full-length product and marker lengths are indicated to the left of the IRD700 image. **(B)** Unpooled individual gel images. Boxed bands represent mutations identified in the pool screen shown in panel A. Other mutations were identified in other pool screens. For each mutation, all eight individual samples comprising a pool were subjected to TILLING analysis after mixing with an equal amount of wild-type DNA. Mutations were confirmed by a complement band in the IRD800 channel.

2.5. CEL I Nuclease Preparation

1. Juicer (e.g., Le Quipe).
2. 0.1 *M* Phenylmethylsulfonyl fluoride (PMSF) (stock in isopropanol); to prepare an aqueous solution of 100 μ*M* PMSF, add 1 mL 0.1 *M* PMSF/L of solution.
3. Buffer A: 0.1 *M* Tris-HCl, pH 7.7, 100 μ*M* PMSF.
4. Buffer B: 0.1 *M* Tris-HCl, pH 7.7, 0.5 *M* KCl.
5. Buffer C: 50 m*M* Tris-HCl, pH 8.0.
6. To prepare ConA-Sepharose® (Bing Yang, personal communication):
   a. Wash 110 mL ConA-Sepharose (Sigma) with 1 L 0.1 *M* HEPES, pH 8.3.
   b. Add 1 g dimethylsuberimidate (Pierce Chemical) dissolved in 0.1 *M* HEPES, pH 8.3. Adjust to pH 8.6 with mixing. React for 1 h at room temperature in a total vol of 500 mL. Add 100 mL of 0.1 *M* glycine, pH 8.3, to neutralize the reaction. Incubate at least 30 min at room temperature.
   c. Recover the resin using a glass-sintered filter with paper. Wash with 200 mL 0.1 *M* NaOAc, 1 *M* NaCl, pH 4.8, then with 200 mL 0.1 *M* Na_2CO_3, 1 *M* NaCl, pH 7.6. Repeat washes two more times. Store in 30% ethanol at 4°C.
7. Nicking buffer: 20 m*M* HEPES, pH 7.5, 10 m*M* KCl, 3 m*M* $MgCl_2$.
8. Stop solution: 50 m*M* Tris-HCl, pH 6.8, 3% sodium dodecyl sulfate (SDS), 4.5% β-mercaptoethanol, 30% glycerol, 0.001% bromophenol blue.

3. Methods

3.1. PCR and Heteroduplex Formation

Perform these steps using equipment and consumables that are segregated from PCR products to avoid contamination. Multipipetors are used to reduce the number of pipeting steps.

1. In advance, array 5 μL of pooled genomic DNA per well in 96-well microtiter plates. These plates can be stored for more than 4 wk at –20°C in a sealed container. The final concentration of genomic DNA and the maximal allowable pooling may vary depending on the organism and DNA extraction method used (*see* **Notes 2–4**). For *Arabidopsis*, the concentration of DNA is 0.015 ng/μL, and samples are pooled eight-fold.
2. Add 5 μL freshly made PCR mixture. Immediately after addition, centrifuge briefly, and place in thermal cycler.
3. Run the following thermal cycler program: 95°C for 2 min; loop 1 for 8 cycles (94°C for 20 s, 73°C for 30 s, reduce temperature 1°C per cycle, ramp to 72°C at 0.5°C/s, 72°C for 1 min); loop 2 for 45 cycles (94°C for 20 s, 65°C for 30 s, ramp to 72°C at 0.5°C/s, 72°C for 1 min); 72°C for 5 min; 99°C for 10 min; loop 3 for 70 cycles (70°C for 20 s, reduce temperature 0.3°C per cycle); hold at 8°C. After thermal cycling store samples in the dark at 4°C for use within 1 wk.

3.2. CEL I Digestion

1. Place PCR samples on ice and add 20 μL CEL I reaction mixture to each sample. Mix by pipeting up and down 2–5 times. The same tips can be reused if rinsed with water between pipeting steps. Incubate at 45°C for 15 min.
2. Place samples on ice and stop reaction by adding 5 μL 0.15 *M* EDTA. Store samples in the dark at 4°C, for use within 1 wk.

3.3. Spin Plate Cleanup

1. Assemble hydrated spin plate and an empty 96-well catch plate for centrifugation.
2. Spin hydrated plate containing Sephadex G50 for 2 min at 440*g*.
3. Remove the catch plate and insert the sample receptacle plate. Load the CEL I digestion products onto the Sephadex plate within 10 min (*see* **Note 5**).
4. Spin for 2 min at 440*g*.
5. Add markers prior to reducing the vol. We add a 200-bp marker onto row D of a 96-well plate and a 95-bp marker in row H (approx 0.5 ng), thus labeling every fourth lane, which also facilitates lane identification (*see* **Subheading 2.3.7.**).
6. Reduce the vol at 85°C to approx 1.5 μL (this takes approx 45 min), leaving formamide–bromphenol blue solution ready for loading. Transfer to ice until ready to load. Samples can be stored in the dark at 4°C for up to 4 wk prior to use.

3.4. IR^2 Gel Analysis

3.4.1. Preparing Gels

1. Assemble 25-cm glass plates, 25-mm spacers, and casting rails. Plates can be preassembled and stored in a dust-free environment for weeks in advance.
2. Pour gels. For each 25-cm plate assembly, fill a 20-mL syringe with freshly prepared acrylamide mixture, then dispense along the top, avoiding bubbles by rapping continuously on the plate just above the liquid edge. If any bubbles appear, remove them quickly after the gel is poured with a thin wire tool.
3. Leaving a little excess acrylamide at the well, insert the top spacer all the way into the glass, making sure spacer is centered horizontally.
4. Insert the Plexiglass pressure plate between the glass plate and casting rails. Tighten the top screws as soon the spacer is inserted, slightly compressing the rubber pads on the pressure plate.
5. Add acrylamide to the top glass edge where the comb is inserted and on the edges to assure that polymerization is not inhibited within the gel.
6. Let the gel set at least 90 min before placing in gel box. Gels can be stored wrapped in plastic wrap at 4°C for up to 24 h prior to use (*see* **Note 6**).
7. Prior to placing gel in gel box, wash the plates with distilled water, removing the comb spacer and excess polyacrylamide at the top edge. Dry the plates and wipe with isopropanol, making sure that the back plate is spotless where the laser shines through.

8. Insert the top buffer reservoir between the glass plate and the casting rails. If this is tricky, moisten the gasket with buffer and remove one casting rail in order to fit the top reservoir. Fill the lower buffer reservoir to the fill line with 0.8× TBE (approx 500 mL), and insert the gel.
9. Tighten the screws to seal the upper reservoir and fill with buffer. Rinse the slot vigorously using a large syringe without a needle (*see* **Note 7**).

3.4.2. Loading Samples onto Membrane Combs

1. Load samples onto a 100-tooth membrane comb, such that position A1 on the plate represents lane 4, position B1 equals lane 5, A2 equals sample 12, etc. (*see* **Note 8**).
2. Use a pipetor to add 0.25–0.5 μL IRD700 plus 800 molecular weight markers to lanes 1, 3, and 100.
3. Spot the IRD700 ladder alone to tooth 2. This asymmetry assures that if the comb is inserted inadvertently in reverse, then the A1 lane is always next to the doubled markers and the H12 lane next to the single marker.

3.4.3. Electrophoresis

1. Access the user controls (LI-COR) using a Netscape browser.
2. Provide a gel run name, hit Create Run (*see* **Notes 9** and **10**).
3. Start the prerun (20 min), waiting for the all ready signal from the scanner before proceeding. The prerun can be started while samples are being applied to the comb.
4. After the prerun, clean the slot out with a syringe and drain the top buffer reservoir until the level is below the glass edge. Wick out the remaining buffer, first with a paper towel and then with a 6-in-wide strip of Whatman 1 paper, sliding it into the slot left by the spacer.
5. Using a 1-mL pipetor, fill the slot with 1% Ficoll® leaving just a thin bead, approx 1 mm above the slot.
6. Hold the comb at a 45°C vertical angle with lane 1 on the left, aim for the slot, and insert rapidly by pushing gently (*see* **Note 11**). Push the comb down until it just touches the gel surface.
7. Gently fill the reservoir to the fill line, insert the electrode–cover, close the top, and then click on Collect image. From the time the comb touches the slot until the time the current is applied should be no more than about 20 s or so to prevent diffusion.
8. After 10 min, open the lid (be sure that you hear the "pling" signal and the high voltage light goes off), remove the comb, and gently rinse the slot with buffer. Replace the top electrode and close the lid. You should hear the pling and see the laser and high voltage lights go on. The gel can be monitored from a browser (*see* **Fig. 2**).

3.5. Gel Image Analysis

Gel images are saved on the LI-COR as tagged image format file (TIFF) images. For visual analysis, the quality of a default JPEG image is sufficient

(*see* **Note 12**). The program "grab" transfers these images to another server via file transfer protocol (ftp) and converts them to JPEG format (*see* **Note 13**). Once grab is done, the layered image can be created.

1. Use Fetch (for a Mac®) or ftp from a Windows® PC to place the two JPEG files onto the desktop or into a local directory.
2. Start up Adobe® Photoshop®, then File > Open the 700 and 800 channel files (**Fig. 2**). Move the 800 channel image to one side, then click on Image > Adjust > Levels, and move the left-most slider arrowhead towards the right until a mid-tone image is obtained for the 800 channel image (usually when the arrowhead is just at the point that the density begins to increase; be sure that Preview is active).
3. Click on OK when the image is optimized. Click on the 700 channel image and repeat the level adjustment procedure. You may want to enlarge the images for setting the levels, which can be done by holding down the Control (or for Mac, the Command) key and press +.
4. Go to Select > All and Edit > Copy, then click on the 800 channel image and Edit > Paste. The 700 image will be precisely superimposed over the 800 image. Close the window to exit from the 700 channel file. You will need the Layers palette, which can be opened by clicking on Window > Show Layers. If your version of PhotoShop does not show rulers on the top and left, click on File > Preferences > Units and Rulers (Edit > Preferences > Units and Rulers on the Mac) and choose percent. If the rulers are not visible, click on View > Show Rulers. Also you can set the grid: File > Preferences > Guides and Grids, choosing "grid line every 5 percent" and 5 subdivisions, as well as a color such as red for the grid lines.
5. Click on Image > Size and change to 2500 (width) × 1750 (height) pixels (uncheck "constrained proportions"), then hold down the Control (or for Mac, the Command) key and press + repeatedly, until the image is at 100%. You can tell if it is, because the ruler will show numbers at 5-U intervals (such as 45....50). Click on View > Fit to screen and adjust the image dimensions as desired (with Image > Size) if needed.
6. Using the Rectangular Marquee Tool, draw a rectangle that encompasses the image from the edge of the electrophoresis front (at bottom) to the full-length product (dark band at top), and from the outside edge of lane 1 to the outside edge of lane 100. Click Image > Crop.
7. In the Layers box, click repeatedly on the eye icon to switch back and forth between the superimposed images. Look for bands present on one image but absent on the other. Ignore bands that are present on both images, as we expect that these are primer dimers resulting from mispriming, not from CEL I cleavage. An exception is a singular position midway in the fragment, where the 700-labeled band coincides in size with the 800-labeled band, and their sum equals the size of the full-length product. In each case, where a band of a specific molecular weight is detected in one channel but absent in the other, look for the complement band in the other channel. Mark each positive lane by double clicking in the vertical ruler area and pulling a vertical line over to the right of the

mutation. Because size is nearly directly proportional to vertical distance, it should be relatively easy to anticipate where to look: for instance, a band one-third of the distance from the front at the bottom of the gel should be paired with a band in the other channel that is one-third of the distance from the full-sized band.

8. Run the program squint (*see* **Note 13**). You will be asked a series of questions about sizes and locations of bands on the image that you are viewing in Photoshop. To ascertain the size corresponding to a band, place the horizontal cross-hair over the band and look at the ruler on the side: the guide line will indicate the distance migrated as a percentage of the distance to the full-sized band. Be sure to compensate for any mobility differences across the gel, such as "frowning" or "smiling," using the background bands as guides. Enter this distance into squint.
9. To determine the lane location, find a favorable vertical position (usually the full-size band) and count to determine the lane position using the markers. Because there are precisely 100 teeth in the comb, the lane number and the horizontal percentage should coincide. When migration distances and lane numbers have been entered, squint returns approximate molecular weights and their sum, which is compared to the molecular weight of the full-length PCR product. If these numbers are nearly equal, then this is almost certainly a mutation. Squint also returns a plate position of the pool given the lane position.

3.6. Analyzing Mutant Individuals

Once squint entries are entered for eight-fold pools, individuals used to make the pools can be efficiently screened to track down each mutation. It is most efficient to screen for individuals once 12 mutations have been entered (if pooling eight-fold), as the screen can then proceed with a full 96-well plate of samples.

1. Run the program pick (*see* **Note 13**). The pick program takes squint output from multiple pool plates and returns a list of rows from plates containing individual DNAs arrayed in an 8 × 8 grid on a 96-well plate. All members of a single pool are present in one row on this plate. These rows will be deposited into successive columns numbered 1–12 on the pick list. This list is the template for the new screening plate to be made containing individuals from the identified mutant pools. Each well in a single column of this plate will contain one individual in the eight-fold pool.
2. From the pick output, take out the individual plates for the first 12 mutations from a single set of oligonucleotides. Rotate the individual plates such that position A1 is in the upper right corner. Using an eight-channel pipetor, transfer 10 μL from each corresponding row of the individual plates to a column in a new plate.
3. Once this plate has been created from 12 rows of individual plates, transfer 5 μL to a new plate containing 5 μL of wild-type DNA (*see* **Note 14**). The original

plate can be stored at –20°C for later amplification and sequencing. TILLING is performed as described in **Subheading 3.1.–3.5.**, and results should resemble that shown in **Fig. 2B**.

3.7. CEL I Preparation

1. This protocol is adapted from refs. ***5*** and ***6***. Celery is available year-round from supermarkets. Rinse and dry to avoid juice dilution by surface water. Cut off the bottom white part and trim the tops if there are many leaves. Juice desired amount of celery (25 kg yields approx 10 L of juice). Adjust the juice to the composition of buffer A (0.1 *M* Tris-HCl, pH 7.7, 100 μ*M* PMSF) with gentle stirring (*see* **Note 15**).
2. Spin the juice in 1-L bottles in a Sorvall® RC-3 (swinging bucket) at 2600*g*, save the supernatant, and discard the pellet. This step is optional, but helps reduce the amount of particulate in the juice.
3. Bring the supernatant to 25% $(NH_4)_2SO_4$ by adding 144 g/L of solution (salt additions will greatly increase vol of the juice). Mix gently at 4°C for at least 30 min. Divide the solution into 500-mL centrifuge bottles and spin in a GSA rotor at 13,000–16,200*g* at 4°C for 40 min. Discard the pellet.
4. Bring the supernatant from 25–80% $(NH_4)_2SO_4$ by adding 390 g/L of solution. Mix gently at least 30 min (or overnight). Spin again in a GSA rotor at 13,000–16,200*g* for at least 1.5 h. Save the pellet and discard the supernatant, being careful in decanting the supernatant as the pellet is very delicate. The pellet can be stored at –80°C.
5. **Step 1** can be repeated and the samples pooled before or after **step 2**.
6. Resuspend the pellets in at least 500 mL buffer B and dialyze thoroughly against buffer B. Do not reduce the resuspension vol of the pellet to <500 mL.
7. Crosslink ConA to Sepharose beads (*see* **Subheading 2.5.6.**). Add 25 mL bed vol of ConA–Sepharose to the sample and gently roll the container overnight at 4°C. Using a glass-sintered filter with paper, filter the resin from the liquid (save the flow-through at –20°C). Wash the resin with buffer B until wash flows through clear (depending on how dirty the pellet was, approx 1–3 L of wash buffer). Elute CEL I from the resin by suspending in 50 mL of buffer B with 0.01% Triton X-100 and 0.3 *M* α-methyl-mannoside, rotate at room temperature for 10 min, then filter out the eluate and save (on ice). Repeat a total of 10× for a final vol of 500 mL eluate. Save the resin at 4°C for reuse.
8. Dialyze the eluae thoroughly against buffer C. Eluate can be frozen at –20°C. After this step and each following step, you can test for CEL I nuclease activity using the plasmid nicking assay (**step 11**).
9. The following is for a 100 mL diethylaminoethyl (DEAE) column. Prewash the column slowly with at least 300 mL of buffer C containing 10 m*M* KCl. Load the sample onto the column with buffer C containing 10 m*M* KCl at 5 mL/min and collect flow-through in fractions. Wash the column with at least 400 mL buffer C containing 10 m*M* KCL. Elute CEL I with a 500-mL gradient from 10 m*M* to 0.5 *M* KCl in buffer C containing 50 m*M* α-methyl-mannoside and collect eluate as

fractions. Step to 1 *M* KCl in buffer C containing 50 m*M* α-methyl-mannoside for 400 mL and collect in 1 vol. Briefly wash with buffer C containing 50 m*M* α-methyl-mannoside. Wash with 500 mL 2 *M* KCl, then water. Finish with a 20% ethanol (filtered) wash for long-term storage. CEL I elutes between 0.1 and 0.2 *M* KCl, with the highest activity at 0.12 *M* KCl. Test fractions for CEL I nicking activity (**step 11**), pool fractions with activity, and dialyze against buffer C. Samples can be stored frozen at –20°C.

10. Equilibrate a Poros HQ column with buffer C at 1.8 mL bed vol, 10 mL/min flow rate, 10-mL fractions, over 15-column vol, in buffer C. Load the supernatant onto a Poros HQ column. Elute CEL I in a linear gradient from 0–1 *M* KCl in buffer C with 50 m*M* α-methyl-mannoside and collect 1-mL fractions. CEL I activity comes off the column between 0.1 and 0.4 *M* KCl. Pool all fractions containing activity, aliquot, and store at –80°C. Test the flow-through for activity using the plasmid nicking assay (**step 11**). The specific activity of CEL I using the nicking assay should be 10^6 U/mL.
11. Plasmid nicking assay for CEL I activity ***(7).*** Incubate 1 μg plasmid with CEL I in 30 μL of nicking buffer for 30 min at 37°C. Add 5 μL stop solution and mix. Electrophorese 24 μL of the final reaction on a 0.8% agarose gel together with an undigested control sample. You should see both shifting of the upper band (covalently closed circles) and streaking of the lower band. A unit of CEL I is defined as the amount of enzyme required to digest 50% of 200 ng of a 500-bp DNA fragment that has a single mismatch in 50% of the duplexes.

4. Notes

1. Special care should be taken when using primers labeled with IRD700 and IRD800. When possible, avoid prolonged exposure of labeled primers and PCR products to fluorescent lights. Primer stocks should be diluted to no more than 100 μM, aliquoted, and stored at –80°C. Primer mixtures are used for no more than about 1 wk. Over time, we see the amount of labeled PCR product decrease dramatically when using old primer mixtures or primer stocks that were stored at 4°C or that have undergone repeated freeze–thaw cycles. We are currently unclear as to the cause of this decrease in efficiency. IRD-tagged oligonucleotides do not prime as well as untagged oligonucleotides, presumably because of the hydrophobic group at the 5′ end. To obtain consistently high PCR product yield, we add a mixture of both tagged and untagged primers. Using the CODDLE program for primer design (http://www.proweb.org/input/), which runs the Whitehead Primer3 program, we have found that >90% of our primer pairs with T_m approx 70°C are successful in amplifying 1-kb fragments from *Arabidopsis* DNA samples and providing adequate TILLING results. Because IRD800 gives a weaker signal than IRD700, it fails more frequently, and mutations might be overlooked when there is only a single channel for detection. This is especially a problem for mutations that are distant from the tagged oligonucleotide priming site, because the large molecular weight strand produced by CEL I digestion has reduced signal and band resolution.

2. The quality and quantification of genomic DNA starting material is crucial. DNA samples for the *Arabidopsis* TILLING Project (http://tilling.fhcrc.org:9366/) are prepared using the FastDNA® kit (Bio101), and samples are electrophoresed on agarose gels to equalize concentrations before arraying and pooling. A single sharp band of high molecular weight indicates high quality DNA.
3. DNA pooling provides higher throughput by allowing less machine time per sample for mutation discovery. For pooling, DNA quantification between samples is very important. As eight-fold pooling approaches the limit of detection for a heterozygous mutation (one-sixteenth), any sample whose concentration is lower than others in the pool may escape mutation detection. Before proceeding with higher pooling and sample prep of thousands of individuals, we suggest trying several different levels of pooling on a small subset of samples. This will also determine the robustness of amplification.
4. Before proceeding with fluorescently tagged primers in PCR, it is desirable to perform trial reactions with unlabeled primers. A yield of 7–10 ng/μL of final PCR product is required for robust and consistent identification of mutations on gel analyzers when pooling samples eight-fold.
5. It is important to deposit sample directly over the center of the Sephadex column, thus avoiding any disturbance of the column, such as touching the column with pipet tips.
6. Gel plates may be prepoured and stored for up to several days at 4°C covered in plastic wrap. Each 25-cm plate requires approx 20 mL. To assemble new plates, clean the plates with dilute liquid detergent (i.e., 2% Tween® 20) and a soft scrub brush. Rinse plates with distilled water, wipe down with 0.2 N HCl, rinse with distilled water again, and wipe with isopropyl alcohol. Spacers are cleaned by wiping with a wet tissue. Assemble the pieces with the screws backed off and align the pieces by standing the assembly vertically while tightening the screws. Tighten just beyond where you begin to feel resistance (over tightening will crack plates). Place on a horizontal support. When handling acrylamide or polyacrylamide gels, always wear gloves, as unpolymerized acrylamide is a nerve toxin.
7. It is important that the slot is clean, as any loose acrylamide will inhibit insertion of the comb. For best visibility of the loading well, insert a background card wedged behind the back plate such that it is centered between the ears of the front plate. Two vertical marks can be made on the card that are a comb width apart and will provide a guide for later inserting the comb precisely in the middle of the gel, which is necessary so that all of the lanes are scanned.
8. A variety of options are available for loading samples. Samples can be loaded onto the membrane comb with the aid of robotics using the comb load robot (MWG). Alternatively, samples can be loaded manually directly onto the comb using a pipetor. We have found that a sample vol of 0.5 μL is optimal for our assays. Manual loading can be aided by using a membrane loading tray (The Gel Company) and an adjustable width multichannel pipetor (Matrix Technologies, Lowell, MA, USA). Combs can be preloaded and stored at 4°C, but as samples are more stable in formamide load buffer, it is suggested that combs be used within

2 h of loading. It is not advised to use loaded combs that have been stored for more than 2 d.

9. Gels can be run twice, even after a day. The prerun is necessary if the plate has been moved, because it is needed for focusing the lasers. After runs are complete, and a new gel is ready, remove the old gel, pour out the buffer from both reservoirs, and clean plates as described in **Subheading 4.6**.
10. For a 1-kb fragment, enter the following settings: collect time 3 h 45 min run at 1500 V, 40 ma, 40 W, 50°C. (Be sure that the current is off before touching a buffer chamber.) Other parameters are pixel size, 16; bin size , 8; and motor speed, 3.
11. Practice this step with used combs. Take special care to avoid bending any teeth. If a tooth sticks to the plate, it may not be possible to save it without moving the comb around and thus losing sample from all teeth.
12. Current versions of the UNIX® programs, grab, squint, and pick are available upon request. The TILLING Web site (http://tilling.fhcrc.org:9366/) provides links to CODDLE, PARSESNP, and SIFT, which facilitate fragment choice, primer design, and mutation analysis.
13. One potential source of error in identifying mutations comes from the misscoring of lanes harboring mutations. Occasionally LI-COR gel images look fuzzy with diffuse and ill-defined lanes, for any of the following reasons:
    a. Urea is not thoroughly rinsed from the well before adding Ficoll.
    b. Comb is pushed into acrylamide when loaded.
    c. Comb moves laterally once inserted into the well.
    d. Upper buffer chamber is filled too rapidly after the comb is inserted.
    e. Ficoll is not rinsed out of well after the comb is removed.
    f. Ficoll is rinsed too thoroughly from the well after the comb is removed.
14. In order to identify individual mutants that are homozygous, samples must be doped with wild-type DNA to generate heteroduplexes that are the substrate for CEL I.
15. All steps should be performed at 4°C, except for the ConA and Poros HQ columns, which can be at room temperature with samples and buffers chilled on ice. CEL I is very stable and dialysis or loading steps can occur overnight if needed. Triton X-100 is essential in later steps, as CEL I tends to aggregate during purification. The DEAE column peak is very broad: you will get about a two-fold purification by collecting only the highest peak, but also throw away about half the protein. Cross-link the ConA-Sepharose on the same day that the celery is juiced. The cross-linking step is performed to reduce contamination from unlinked ConA.

Acknowledgments

The methods described here were developed with support from the National Science Foundation Plant Genome Research Program. We thank Chris Burtner and Anthony Odden for the images shown in Figure 2, Bing Yang for the ConA-Sepharose protocol, and Tony Yeung for advice on CEL I preparation and use.

References

1. McCallum, C. M., Comai, L., Greene, E. A., and Henikoff, S. (2000) Targeted screening for induced mutations. *Nat. Biotechnol.* **18,** 455–457.
2. McCallum, C. M., Comai, L., Greene, E. A., and Henikoff, S. (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol.* **123,** 439–442.
3. Colbert, T., Till, B. J., Tompa, R., Reynolds, S., et al. (2001) High-throughput screening for induced point mutations. *Plant Physiol.* **126,** 480–484.
4. Middendorf, L. R., Bruce, J. C., Bruce, R. C., et al. (1992) Continuous on-line DNA sequencing using a versatile infrared laser scanner/electrophoresis apparatus. *Electrophoresis* **13,** 487–494.
5. Oleykowski, C. A., Bronson Mullins, C. R., Godwin, A. K., and Yeung, A. T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Res.* **26,** 4597–4602.
6. Yang, B., Wen, X., Kodali, N. S., et al. (2000) Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry* **39,** 3533–3541.
7. Yeung, A. T., Mattes, W. B., Oh, E. Y., and Grossman, L. (1983) Enzymatic properties of purified *Escherichia coli uvrABC* proteins. *Proc. Natl. Acad. Sci. USA* **80,** 6157–6161.

14

Gene and Enhancer Traps for Gene Discovery

Marcela Rojas-Pierce and Patricia S. Springer

Summary

Gene traps and enhancer traps provide a valuable tool for gene discovery. With this system, genes can be identified based solely on the expression pattern of an inserted reporter gene. The use of a reporter gene, such as β-glucuoronidase (*GUS*), provides a very sensitive assay for the identification of tissue- and cell-type specific expression patterns. In this chapter, protocols for examining and documenting *GUS* reporter gene activity in individual lines are described. Methods for the amplification of sequences flanking transposant insertions and subsequent molecular and genetic characterization of individual insertions are provided.

Key Words

Arabidopsis, *Ds*, gene trap, enhancer trap, transposable element, GUS, gene expression, TAIL-PCR, transposon tagging, mutagenesis

1. Introduction

Gene and enhancer trap insertions allow the identification of genes based on the expression pattern of a reporter gene. The basic principle relies on the creation of random genomic insertions of a reporter gene, the expression of which can be easily visualized. When the reporter gene is inserted within or nearby to a chromosomal gene, the reporter gene becomes regulated by the chromosomal gene. Reporter gene expression in specific tissues or cell types can be identified, in addition to those that show temporal or conditional regulation of expression. This approach has been used successfully in a number of different organisms, including bacteria, *Drosophila*, mice, and plants ***(1–4)***. In plants, most trapping systems have been developed to take advantage of transposons or T-DNA as the insertional agent and *β-glucuronidase* (*GUS*) as the reporter gene of choice. The *green fluorescent protein* (*GFP*) and selectable marker

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

genes have also been utilized as reporter genes (reviewed in ref. ***3***). Many gene-trap systems have been developed in the model plant *Arabidopsis thaliana*. Recently, gene-trap systems have become available in rice, *Lotus japonicus*, and *Physcomitrella patens* ***(5–8)***.

We describe here methods for GUS staining and molecular and genetic characterization of gene and enhancer trap transposants that are based on the transposable element *Ds*. Many collections of gene and enhancer trap lines have been generated (reviewed in ref. ***3***), and seed from these collections is available from the Arabidopsis Biological Resource Center (*see* Chapter 19 by R. Scholl et al., this text). We begin our discussion with the identification of reporter gene expression patterns in individual lines using GUS staining. Detailed protocols for the generation of transposants using the Cold Spring Harbor gene-trap system have been described elsewhere ***(9)***. We have focused on the most common situations within each section to maintain clarity, but variations may occur in each case depending of the genetic characteristics of the insertion. Many of the protocols described here can be modified for use with systems other than the Cold Spring Harbor one. An excellent reference for general protocols in *Arabidopsis* can be found in the recently published *Arabidopsis* protocols manual by Weigel and Glazebrook ***(10)***.

2. Materials

2.1. Screening for GUS Expression Patterns

2.1.1. GUS Staining

1. GUS stain solution: 100 m*M* sodium phosphate buffer, pH 7.0 (from 1 *M* stock, *see* **step 2**), 10 m*M* ethylene diamine tetraacetic acid (EDTA), 0.1% Triton® X-100, 1 mg/mL 5-bromo-4-chloro-3-inolyl-β-D-glucuronic acid, cyclohexyl-ammonium salt (X-Gluc; Biosynth International, cat. no. B-7300), 100 μg/mL chloramphenicol, 2 m*M* potassium ferricyanide, 2 m*M* potassium ferrocyanide (*see* **Note 1**). X-Gluc will dissolve in an aqueous solution very slowly; therefore it is convenient to first dissolve the X-Gluc in dimethylformamide (DMF) at 100 μg/mL immediately before use.
2. 1 *M* sodium phosphate buffer, pH 7.0: 57.7 mL 1 *M* Na_2HPO_4, 42.3 mL 1 *M* NaH_2PO_4.
3. 24- or 48-well tissue culture plates (Corning Costar, cat. no. 3524 or 3548).
4. Vacuum desiccator.
5. 70% Ethanol.
6. 50% (v/v) Glycerol.
7. Stereo microscope with light source.
8. Compound microscope equipped with dark field illumination and differential interference contrast (DIC) optics.

2.1.2. Photography

1. Stereo microscope with fiber optic gooseneck light source.
2. Compound microscope equipped with dark field illumination and DIC optics.
3. 35 mm or digital camera.
4. Glass slides.
5. Cover glasses, no. 1.5.
6. 50% glycerol.

2.1.3. Sectioning of GUS-Stained Tissue

1. 100 m*M* sodium phosphate buffer, pH 7.0 (made from 1 *M* stock, *see* **Subheading 2.1.1.**).
2. Glutaraldehyde fixative: 2.5% glutaraldehyde in 100 m*M* sodium phosphate buffer, pH 7.0, diluted from 25% electron microscopy grade glutaraldehyde (EM Science, cat. no. 16216). Prepare fresh and keep on ice.
3. Phosphate-buffered saline (PBS): 137 m*M* NaCl, 2.7 m*M* KCl, 10 m*M* Na_2HPO_4, 2 m*M* KH_2PO_4, adjust pH to 7.4 with HCl. Can be made as a 10× stock.
4. Ethanol: 100%, 95%, 80%, 70%, 50%, and 30%.
5. Paraplast Plus (Fisher Scientific, cat. no. 23–021400).
6. 60°C Oven.
7. Glass scintillation vials.
8. CitriSolv citrus clearing agent (Fisher Scientific; cat. no. 22–143–975).
9. Rotary microtome.
10. Coated microscope slides (Fisherbrand Superfrost Plus; Fisher Scientific; cat. no. 12–550–15).
11. Variable temperature hot plate.
12. Mount-Quick aqueous mounting media (EM Science; cat. no. 18002).

2.2. Amplification of DNA Sequences Flanking Ds Insertion

2.2.1. DNA Isolation

1. Kontes disposable pellet pestle (Fisher Scientific; cat. no. K749521–1590).
2. Urea extraction buffer: 420 g/L urea, 0.3 *M* NaCl, 50 m*M* Tris-HCl, pH 8.0, 20 m*M* EDTA, pH 8.0, 1% (w/v) sarkosyl (*N*-lauroylsarcosine).
3. Phenol:chloroform:isoamylalcohol, 25:24:1, prepared with phenol buffered against Tris-EDTA (TE).
4. Isopropanol.
5. 3 *M* Sodium acetate, pH 5.2.
6. TE: 10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA.
7. 100% and 70% Ethanol.

2.2.2. Thermal Asymmetric Interlaced Polymerase Chain Reaction

1. *Ds*-specific primers ***(11)***, 2 μ*M* stocks.
   Ds3-1 5'-ACCCGACCGGATCGTATCGGT-3'.
   Ds3-2 5'-CGATTACCGTATTTATCCCGTTC-3'.
   Ds3-4 5'-CCGTCCCGCAAGTTAAATATG-3'.
   Ds5-1 5'-ACGGTCGGGAAACTAGCTCTAC-3'.
   Ds5-2 5'-CCGTTTTGTATATCCCGTTTCCGT-3'.
   Ds5-4 5'-TACGATAACGGTCGGTACGG-3'.
2. Arbitrary degenerate primers for thermal asymmetric interlaced polymerase chain reaction (TAIL-PCR), 20 μ*M* stocks.
   AD1 5'-NTCGA(G/C)T(A/T)T(G/C)G(A/T)GTT-3' ***(12)***.
   AD2 5'-NGTCGA(G/C)(A/T)GANA(A/T)GAA-3' ***(12)***.
   AD5 5'-(A/T)CAGNTG(A/T)TNGTNCTG-3' ***(11)***.
3. *Taq* DNA polymerase, 5 U/μL (Fisher Scientific).
4. 10× PCR buffer A (supplied with *Taq* DNA polymerase): 500 m*M* KCl, 15 m*M* $MgCl_2$, 100 m*M* Tris-HCl, pH 9.0.
5. Deoxyribonucleotide (dNTP) mixture containing 2 m*M* each of dATP, dTTP, dGTP, and dCTP.

2.3. Sequencing of TAIL-PCR Products

1. QIAquick™ PCR purification kit (Qiagen, cat. no. 28104).
2. Ds3-4 or Ds5-4 primer, 5 μ*M* (**Subheading 2.2.2**).

2.4. Southern Blot Analysis of Transposant Line

1. *Eco*RI (Promega).
2. 10× Buffer H (supplied with enzyme).
3. Bovine serum albumin (BSA) 1 mg/mL.
4. Spermidine (100 m*M*, pH 7.0).
5. Nylon membrane (Osmonics; Fisher Scientific; cat. no. N00HY00010).
6. Church hybridization buffer ***(13)***: 1% (w/v) BSA, 1 m*M* EDTA, 0.25 *M* phosphate buffer (from 1 *M* stock, *see* **step 7**), 7% (w/v) sodium dodecyl sulfate (SDS).
7. 1 *M* Phosphate buffer: dissolve 142 g Na_2HPO_4 in 800 mL of water, adjust pH to 7.2 with 85% H_3PO_4, and bring to final vol of 1 L.
8. 20× SSC: 175.3 g/L NaCl, 88.2 g/L sodium citrate. Adjust pH to 7.0 with 14 N HCl.
9. 20% SDS.
10. 2× SSC, 0.2% SDS (made from 20× SSC and 20% SDS).
11. 0.2× SSC, 0.2% SDS (made from 20× SSC and 20% SDS).
12. Kodak® X-Omat™ film (Eastman Kodak).

2.5. Sequence Analysis

1. Computer with access to the Internet.

3. Methods

3.1. Screening for GUS Expression Patterns

3.1.1. GUS Staining

Histochemical detection of GUS activity in most *Arabidopsis* tissues is very straightforward. X-Gluc is expensive, so care should be taken to minimize the vol of stain solution. Use just enough to cover the tissue. Many tissues, including intact seedlings, can be stained in 24- or 48-well tissue culture plates. If a conditional screen to identify lines that show a change in *GUS* expression under a particular environmental condition is to be carried out, carefully designed controls are extremely important. In particular, the developmental stage of control plants and treated plants must be matched.

1. Harvest the tissue of interest directly into GUS stain solution, making sure it is entirely submerged. Do not let the tissue dehydrate or the stain may not penetrate well. In many cases, the gene-trap insertion will be segregating in families that are to be screened for GUS activity. In order to be certain that plants containing the insertion are sampled, 4–6 plants from each family should be stained.
2. Place the tissue culture plate in a vacuum desiccator and draw a vacuum (house vacuum is normally sufficient) for 15 min.
3. Release the vacuum, seal the plates with parafilm or tape, and wrap them in aluminum foil to exclude light.
4. Incubate the staining reaction at 37°C for 24–48 h. For initial screens, we recommend incubating for 48 h to allow detection of GUS activity in lines that have low levels of *GUS* expression. The long incubation may result in overstaining and decreased specificity in lines showing high levels of *GUS* expression however. Lines of interest can be examined in a secondary screen using shorter incubation times. In each experiment, include a control line with a characterized *GUS* expression pattern.
5. Remove the GUS stain solution and replace it with 70% ethanol to remove chlorophyll from the tissue.
6. Incubate in 70% ethanol at room temperature, changing the ethanol when it turns green, as many times as necessary. It is important to adequately clear chlorophyll from the tissue in order to visualize faint *GUS* expression patterns. This can take 1–3 d, depending on the tissue. Clearing will proceed more rapidly at 37°C, but do not to let the ethanol evaporate. The stained tissue can be stored in 70% ethanol at 4°C for several wk, but it will become more fragile with time. Evaporation of the ethanol causes internal air bubbles to develop in the tissue, so store the plates in an airtight container if possible.
7. Examine the tissue for *GUS* expression patterns. This is often the most time-consuming step in a large-scale screen. Be sure to allow adequate time for careful examination. It is convenient to perform the initial screen for GUS activity in the tissue culture plates using a stereomicroscope. With practice, even faint expres-

sion patterns will be detectable. As needed, stained tissue can be placed in 70% ethanol in a Petri dish for manipulation and closer examination. Tissue can also be mounted on a glass slide in 50% glycerol and viewed using a compound microscope with DIC optics. In most cases, the tissue can be transferred directly from 70% ethanol to 50% glycerol without a substantial loss of integrity, and rehydration is not necessary.

3.1.2. Photography

It can be challenging to obtain good images of GUS-stained tissue, especially in the case of very faint expression patterns. Initially, you will probably wish to photograph tissue using a stereomicroscope. In this case, background and lighting conditions are the most important factors in obtaining high-quality images. We have had reasonable success using agar media as a background. A fiber optic gooseneck light source should be used if possible, as this allows the most control over lighting angle.

1. Float the tissue in 50% glycerol on an agar plate (made with water) and position as needed.
2. Position the light source to minimize glare. Indirect lighting from the side often works well.
3. Remove as much liquid as possible and photograph. The agar media will prevent the tissue from drying out if you work quickly. If using traditional 35 mm photography, be sure to bracket the exposure times broadly.

   To obtain images at higher magnifications, the tissue can be mounted on a glass slide under a cover glass.
4. Place tissue on a slide and remove as much ethanol as possible.
5. Place a drop of 50% glycerol over the tissue.
6. Slowly place a cover glass over the tissue, being careful not to introduce air bubbles.
7. Remove excess glycerol by wicking it away from the side with filter paper.
8. View the tissue using DIC optics and photograph. Faint GUS staining will be more easily visible under dark field illumination, which results in the stained tissue appearing pink.

3.1.3. Sectioning GUS-Stained Tissue

Higher resolution visualization of *GUS* expression patterns can be obtained in tissue sections. In some cases, hand sections of fresh tissue can be used, but this is not satisfactory in many instances. Embedding in wax and sectioning is an easy alternative. The following protocol is modified from *ref.* ***10***.

1. Wash the GUS-stained tissue in 100 m*M* sodium phosphate buffer, pH 7.0, 1 h to overnight.
2. Replace the phosphate buffer with glutaraldehyde fixative and vacuum infiltrate the tissue on ice. Glutaraldehyde must be handled in a chemical fume hood while

wearing appropriate gloves and goggles because it is highly toxic if it is inhaled or comes in contact with skin.
3. Allow tissue to fix for 60 min.
4. Wash with ice-cold PBS, 3 × 10 min.
5. Dehydrate the tissue through an ethanol series (30%, 50%, 70%, 80%, 95%, and 100%), 90 min each step at 4°C.
6. Change the 100% ethanol and leave at 4°C overnight.
7. Perform two additional incubations in 100% ethanol, 2 × 60 min each at room temperature to be sure that no water remains.
8. Replace the ethanol with CitriSolv (do not use xylene, it will dissolve the indigo precipitate) using a series of CitriSolv/ethanol mixtures at room temperature: 25% CitriSolv/75% ethanol for 60 min, 50% CitriSolv/50% ethanol for 60 min, 75% CitriSolv/25% ethanol for 60 min, 100% CitriSolv 3 × 60 min.
9. Replace CitriSolv and add 25% vol of Paraplast wax chips, incubate at 42°C overnight.
10. Add another 25% vol of Paraplast chips and move to 60°C.
11. Replace wax/CitriSolv mixture with freshly melted wax after the chips have melted.
12. Replace with freshly melted wax twice a day for 3 d. The Paraplast temperature should not exceed 60°C, and melted wax should not be stored longer than 1 wk. Work quickly when changing wax, as it will solidify rapidly.
13. Pour tissue into molds on a variable temperature hot plate and position using hot needles or forceps.
14. Gradually move the mold to cooler positions on the plate.
15. When the wax hardens, cut blocks that contain the tissue, mount, trim, and section using a rotary microtome. For faint expression patterns, you may want to cut fairly thick sections (>10 μm).
16. Float the sections on a slide in water at 42°C on a slide warmer.
17. When the sections have expanded, remove the water and dry on the slide warmer overnight.
18. Deparaffinize them by soaking in CitriSolve, 2 × 10 min.
19. Rehydrate the tissue through a series of ethanol steps: 100%, 95%, 80%, 70%, 50%, and 30%, 10 min each, then transfer to water.
20. Mount under a cover glass in aqueous mounting media.
21. Allow the mounting media to dry overnight, and observe with the microscope using DIC optics.

3.2. Amplification of DNA Sequences Flanking Ds *Insertion*

3.2.1. DNA Isolation (14)

Genomic DNA from wild-type Landsberg *erecta* (*Ler*) and the transposant lines of interest can be isolated from a few inflorescences or 8–10 seedlings.

1. Harvest tissue, pooling at least four different transposant plants if they were not selected on kanamycin.

2. Grind the tissue in a microcentrifuge tube in 600 μL of urea extraction buffer using a disposable pellet pestle.
3. Shake for 5 min at room temperature.
4. Add 500 μL phenol:cholorform:isoamylalcohol (25:24:1). Phenol and chloroform are extreme irritants and may be carcinogenic, so wear appropriate chemical resistant gloves and goggles and work in a chemical fume hood.
5. Shake at room temperature for 10 min.
6. Spin at top speed in a microcentrifuge for 5 min.
7. Transfer the aqueous (top) phase to new microcentrifuge tube and add 50 μL of 3 M sodium acetate, pH 5.2, and 500 μL of isopropanol to precipitate nucleic acids.
8. Spin for 5 min in a microcentrifuge.
9. Remove the supernatant and resuspend the resulting pellet in 500 μL of TE.
10. Precipitate the DNA a second time by adding 1 mL of 100% ethanol.
11. Mix and spin in a microcentrifuge to pellet the DNA.
12. Remove the supernatant and rinse the pellet with 70% ethanol.
13. Spin for 5 min in a microcentrifuge.
14. Remove the 70% ethanol and allow pellet to air-dry for about 30 min.
15. Resuspend the pellet in 35 μL TE.
16. Examine the DNA quality and quantity by running 3 μL of each DNA sample on a 0.6% agarose gel. To estimate the DNA concentration, compare the band intensity to that of several different quantities of DNA of known concentration, such as uncut bacteriophage λ DNA.

3.2.2. TAIL-PCR

Amplification of genomic sequences flanking the *Ds* element can be performed by TAIL-PCR ***(11,12)*** (*see* Chapter 15 by T. Singer and E. Burke, this text). Three successive PCR reactions are performed, using arbitrary degenerate (AD) primers and nested primers specific to one end of the *Ds* element. The primers Ds3-1, Ds3-2, and Ds3-4 are nested primers for the 3' end of *Ds*, and Ds5-1, Ds5-2, and Ds5-4 are nested primers for the 5' end. Ds3-4 and Ds5-4 are the most distal in each case ***(11)*** (*see* **Fig. 1**).

1. Set up six independent primary TAIL-PCRs using the six primer combinations listed in **Table 1** (*see* **Note 2**). When amplifying more than one template, make a "cocktail" that contains all common ingredients. Not all primer combinations will result in a successful PCR amplification for each insertion. To increase the probability of successful amplification of DNA flanking each insertion, it is recommended that all six combinations be used. Each primary reaction contains 2 μL of 10× PCR buffer, 2 μL of 2 m*M* dNTP mixture, 4 μL of the appropriate AD primer stock (AD1, AD2, or AD5 at 20 μ*M*), 2 μL of the appropriate *Ds* primer stock (Ds3-1 or Ds5-1 at 2 μ*M*), 10-20 ng of genomic DNA template, and 1 U of *Taq* DNA polymerase in a total vol of 20 μL. Use the PCR conditions shown in **Table 2** ***(11)*****.**

Table 1
Primer Combinations for TAIL-PCR Amplification of *Ds*-Flanking Sequences

	Primary	Secondary	Tertiary	*Ds* end amplified	Sequencing primer
Reaction 1	AD1, Ds5-1	AD1, Ds5-2	AD1, Ds5-4	5'	Ds5-4
Reaction 2	AD1, Ds3-1	AD1, Ds3-2	AD1, Ds3-4	3'	Ds3-4
Reaction 3	AD2, Ds5-1	AD2, Ds5-2	AD2, Ds5-4	5'	Ds5-4
Reaction 4	AD2, Ds3-1	AD2, Ds3-2	AD2, Ds3-4	3'	Ds3-4
Reaction 5	AD5, Ds5-1	AD5, Ds5-2	AD5, Ds5-4	5'	Ds5-4
Reaction 6	AD5, Ds3-1	AD5, Ds3-2	AD5, Ds3-4	3'	Ds3-4

2. Dilute the products of the primary TAIL-PCR 50-fold, and set up the following secondary TAIL-PCR using the primer combinations shown in **Table 1**. The secondary reaction contains 2 μL of 10× PCR buffer, 2 μL of 2 m*M* dNTP mixture, 2 μL of the appropriate AD primer stock (AD1, AD2, or AD5 at 20 μ*M*), 2 μL of the appropriate *Ds* primer stock (Ds3-2 or Ds5-2 at 2 μ*M*), 1 μL of diluted primary PCR products, and 1 U of *Taq* DNA polymerase in a total vol of 20 μL. Use the PCR conditions shown in **Table 2** ***(11)***.
3. Dilute the products of the secondary TAIL-PCR 20-fold, and set up the following tertiary TAIL-PCR using the primer combinations shown in **Table 1**. Each tertiary reaction contains 2 μL of 10× PCR buffer, 2 μL of 2 m*M* dNTP mixture, 2 μL of the appropriate AD primer stock (AD1, AD2, or AD5 at 20 μ*M*), 2 μL of the appropriate *Ds* primer stock (Ds3-4 or Ds5-4 at 2 μ*M*), 1 μL of the diluted secondary TAIL-PCR products, and 1 U of *Taq* DNA polymerase in a total vol of 20 μL. Use the PCR conditions shown in **Table 2** ***(11)***.
4. Check the success of each reaction by running 10 μL of the secondary and tertiary PCR products in adjacent lanes of a 1.5% agarose gel. Successful amplification will result in the presence of visible products in both reactions, with a characteristic size shift between the secondary and tertiary products of 55 bp for the 3' end and 86 bp for the 5' end (*see* **Note 3**).

3.3. Sequencing of TAIL-PCR Products

1. Purify tertiary PCR products from successful TAIL-PCRs using a QIAquick PCR purification kit or similar kit from another manufacturer. Follow the manufacturer's instructions. It may be necessary to repeat the tertiary TAIL-PCRs to obtain a sufficient amount of PCR product for sequencing.
2. Determine the sequence of each tertiary TAIL-PCR product with the appropriate *Ds* primer (Ds3-4 or Ds5-4) (**Table 1**) (*see* **Note 4**).

3.4. Southern Blot Analysis of Transposant Line

It is important to confirm that the amplified and sequenced TAIL-PCR product represents genomic DNA flanking the insertion of interest. The nature of

Table 2
Conditions for TAIL-PCRs *(11)*

Primary			Secondary			Tertiary		
Temp	Time	No. of cycles	Temp	Time	No. of cycles	Temp	Time	No. of cycles
93°C	2 min	1	93°C	1 min	1	93°C	1 min	1
94°C	1 min	5	94°C	30 s	13	94°C	30 s	20
62°C	1 min		62°C	1 min		45°C	1 min	
72°C	2 min		72°C	2 min		72°C	2 min	
			94°C	30 s				
94°C	1 min	1	62°C	1 min		72°C	5 min	1
ramp	3 min		72°C	2 min				
to 25°C	(0.4°C/s)		94°C	30 s		4°C	hold	
25°C	3 min		45°C	1 min				
ramp	3 min		72°C	2 min				
to 72°C	(0.3°C/s)							
72°C	2 min		72°C	5 min	1			
94°C	30 s	15	4°C	hold				
65°C	1 min							
72°C	2 min							
94°C	30 s							
65°C	1 min							
72°C	2 min							
94°C	30 s							
45°C	1 min							
72°C	2 min							
72°C	5 min	1						
4°C	hold							

the PCR technique allows for amplification of extremely small amounts of template. This means that there is the possibility of amplification of contaminating sequences, rather than the sequence of interest. In addition, a small percentage of transposant lines contain more than one *Ds* insertion ***(15)***. In this case, it is important to determine which *Ds* element corresponds to the *GUS* expression pattern of interest and, furthermore, which TAIL-PCR product corresponds to the insertion of interest.

Southern blot hybridization is used to confirm the authenticity of each TAIL-PCR product and to determine the number of *Ds* insertions present. See **Fig. 1** for the location of the *Eco*RI restriction sites.

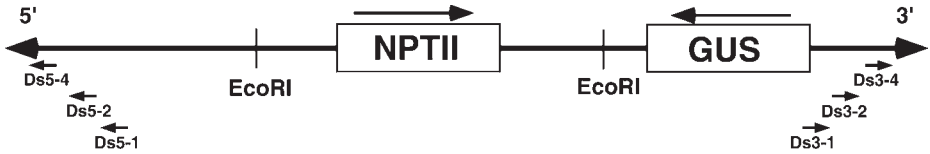


Fig. 1. Schematic showing the *Ds* element used in Cold Spring Harbor enhancer and gene traps. Both the enhancer trap *DsE* and gene trap *DsG* elements have the same general structure, differing only in the sequences upstream of the *GUS* gene at the 3' end of the elements ***(33)***. The locations of the primers used for TAIL-PCR (Ds5-1, Ds5-2, Ds5-4, Ds3-1, Ds3-2, Ds3-4) are shown by arrows (not to scale). The positions of the two *Eco*RI sites in the element are shown. The *NPTII* gene, conferring kanamycin resistance, and the *GUS* gene are shown, with arrows indicating the direction of transcription.

3.4.1. Determining the Veracity of the TAIL-PCR Product

1. Digest genomic DNA isolated from wild type *Ler* and from the transposant line of interest in the following reaction. Mix 5 μg of genomic DNA with 2 μL of 10× buffer H, 2 μL BSA (1 mg/mL), 0.5 μL spermidine (100 m*M*), and 1 μL *Eco*RI (12 U) in a total vol of 20 μL.
2. Incubate at 37°C overnight.
3. Separate the digestion products through a 1% agarose gel, transfer to a nylon membrane, and UV cross-link ***(16)***.
4. Hybridize the membrane with a ^{32}P-labeled TAIL-PCR product. Hybridizations can be performed in Church buffer ***(13)*** at 65°C overnight.
5. Wash the membrane at 65°C in 2× SSC, 0.2% SDS for 20 min, followed by two washes in 0.2× SSC, 0.2% SDS for 20 min each.
6. Expose the membrane to X-ray film for an appropriate length of time, typically 24–48 h.
7. Develop the autoradiograph and evaluate. A polymorphism due to insertion of the *Ds* element should be apparent between wild-type and the transposant line if the TAIL-PCR product represents DNA flanking a *Ds* element.

3.4.2. Determining the Ds Element-Copy Number

To determine the number of *Ds* element insertions in the transposant line, the same membrane is hybridized with a probe for the *Ds* element.

1. Strip the membrane to remove hybridized probe by incubating in 1× TE. buffer at 95°C for 15 min.
2. Hybridize the membrane with a ^{32}P-labeled probe isolated from the *GUS* gene, wash, and expose to X-ray film as described in **Subheading 3.4.1**, **step 4–6**.
3. Develop the autoradiograph and evaluate. If the transposant line contains a single *Ds* element, only one band will be detected by hybridization with the *GUS* probe.

If more than one *Ds* insertion is present, it is possible to isolate individuals containing single elements from a segregating population, unless the elements are closely linked. The majority of lines with multiple elements have only two insertions, and that is the situation we will consider here. In rare instances, more than two insertions are present. In this case, the same approaches can be used, but a larger number of individuals must be examined.

You may already have a population that is segregating for both *Ds* insertions, in which case it will be possible to identify individual plants within that population that have only one of the two elements.

4. To determine if the insertions are segregating, examine the ratio of kanamycin resistant to kanamycin sensitive plants in the population by plating the seed on MS-agar plates supplemented with 50 μg/mL of kanamycin. In the case of two unlinked *Ds* elements, one-sixteenth of the seedlings will be kanamycin-sensitive if both insertions were heterozygous in the parental plant. If either insertion was homozygous, all progeny will be kanamycin-resistant. If one or both of the insertions is homozygous, you will need to perform an outcross to wild-type, followed by a self to generate a segregating population.
5. Grow 16–20 individual plants from a segregating population, isolate genomic DNA from each plant. Allow plants to self-pollinate and harvest seed.
6. Determine the number of *Ds* elements present in each plant, using the Southern blot analysis described above.
7. Repeat the GUS staining on individual plants that have each distinct insertion to determine which element confers the *GUS* expression pattern of interest.
8. Perform the TAIL-PCR amplification on a genomic DNA template that was isolated from a plant with a single insertion of the element of interest.

3.5. Sequence Analysis

Once you have determined that the TAIL-PCR fragment corresponds to the site of *Ds* insertion, the sequence can be used to determine the genomic location of the *Ds* insertion.

1. Examine the sequence that you obtained from the TAIL-PCR product. Identify the part of the sequence that corresponds to the end of the *Ds* element and the sequences corresponding to the cloning vector and trim.
2. Use the remaining sequence as a query in a Basic Local Alignment Search Tool (BLASTN) search against the *Arabidopsis* genome (http://www.Arabidopsis.org/Blast/index.html) to determine the chromosomal location of the *Ds* insertion.
3. Determine the orientation of the *GUS* gene within the chromosomal sequence. *GUS* is oriented so that it is transcribed inward from the 3' end of the *Ds* element.
4. Examine the annotated genome sequence to identify known or hypothetical genes that lie in the vicinity of the insertion site. In the case of a *DsG* gene-trap transposant, the insertion should be within a transcribed gene and oriented so that

GUS and the endogenous gene are transcribed in the same orientation (*see* **Note 5**). Because *GUS* expression from an enhancer trap insertion does not rely on the formation of a transcriptional fusion, *DsE* enhancer trap elements might be found some distance from the endogenous gene that regulates *GUS* expression. The annotation of the *Arabidopsis* genome is continuously updated, and current information can be found at the following Web sites: (www.tigr.org/tdb/e2k1/ath1/) and (http://mips.gsf.de/proj/thal/db/index.html).

5. If no annotated genes are reported in the vicinity of the insertion, several options can be considered. The insertion may lie within or nearby to a gene that has not been predicted by annotation. In particular, genes encoding small peptides and noncoding RNAs, such as microRNAs, are likely to be underrepresented in the current annotation ***(17–22)***. It may be possible to identify genes that have not yet been annotated by probing Northern blots and cDNA libraries with genomic DNA probes from near the site of insertion. The genomic DNA probes can be generated by PCR amplification. It is also possible that *GUS* expression is regulated by a gene that is located a significant distance from the insertion site (*see* **Note 6**).

3.6. Determine Whether GUS Expression Accurately Reflects the Expression of the Tagged Gene

After a candidate gene has been identified, its native expression pattern should be examined and compared to that of the *GUS* expression pattern in the transposant line.

1. Examine the expression of the native gene in wild-type plants using Northern blot analysis, reverse transcription polymerase chain reaction (RT-PCR), or *in situ* hybridization. The choice of technique will depend on the distribution and apparent level of expression, as based on GUS activity. Northern blots and RT-PCR can give some indication of transcript distribution and abundance, but do not provide information about the distribution of transcripts within a tissue or organ. *In situ* hybridization provides the best confirmation of the expression pattern of the endogenous gene, but may not be sensitive enough to allow detection of the very low levels of expression that can be reported by *GUS*, which is extremely sensitive ***(23)***. Therefore, it may not always be possible to detect transcripts using *in situ* hybridization (*see* **Note 7**).

It is also advisable to examine the expression pattern of the endogenous gene using promoter–reporter gene fusions.

2. Design primers to PCR amplify genomic sequences upstream of the translation start site. Regulatory sequences appear to be quite variable in size, so we recommend using genomic sequences that extend as far as the next upstream gene. In the majority of cases, 5' sequences are sufficient for proper regulation, however in some cases, regulatory sequences reside 3' to the transcribed region or in introns or exons within the gene (e.g., ***24–26***).
3. Clone the promoter into a binary vector such as pCB307 or pCB308 ***(27)***, which contain *GFP* and *GUS* reporter genes, respectively (*see* **Note 8**).
4. Introduce the binary vector into *Agrobacterium tumefaciens* strain GV3101 ***(28)***.
5. Transform the T-DNA construct into wild-type *Arabidopsis* plants using the flo-

ral dip method ***(29)***. Multiple transgenic lines containing independent promoter–reporter gene insertions should be generated and examined, as the genomic location of the transgene can result in variability in expression.

3.7. Determine Whether There Is a Mutant Phenotype Associated with the Insertion

In addition to reporting the expression of adjacent chromosomal genes, a transposant insertion might also result in gene disruption. For this reason, the transposant line of interest should be examined for the presence of a mutant phenotype.

1. Grow plants that are homozygous for the insertion of interest and examine them for phenotypic abnormalities.
2. Phenotypes may be uncovered by subjecting plants to a variety of different growth conditions, such as high and low temperatures, treatment with exogenous hormones, etc.
3. In many cases, the expression pattern can serve as a guide in the search for a phenotype. For example, a gene-trap insertion in the *ROP-GAP4* gene was identified by the up-regulation of *GUS* expression during exposure to anoxia ***(30)***. Under normal growth conditions, no phenotype was visible in homozygous plants carrying the *DsG* insertion in *ROP-GAP4,* however these plants exhibited a decreased tolerance to low oxygen stress ***(30)***.
4. If a mutant phenotype is identified, it is important to verify that it is caused by the insertion of interest, as untagged mutations are present in most T-DNA and transposon populations ***(15,31–33)***. A number of different approaches can be used to determine whether a mutation is caused by a transposon insertion (*see* **Note 9**). The most direct way of demonstrating that disruption of a gene is responsible for the observed phenotype is to complement the mutation by introduction of a wild-type copy of the gene.
5. Ideally, complementation will be carried out by introduction of a full-length cDNA clone or a genomic clone that spans the entire gene. In either case, it is preferable to use the gene's native promoter to drive expression, but a strong ubiquitously expressed promoter, such as the cauliflower mosaic virus *35S* promoter ***(34)***, may also be suitable. Use of a promoter such as *35S* is likely to result in expression at higher than normal levels and in ectopic expression. While this may allow complementation of the mutant phenotype, other phenotypic consequences are likely. If the mutation is recessive, and homozygous plants are fertile, then the complementation construct may be directly transformed into the mutant background. In this case, primary transformants can be directly examined for a rescue of the mutant phenotype. In the case of a mutation that causes lethality or decreased fertility, wild-type plants can be transformed, and the complementation construct can later be introduced into the mutant background by genetic crossing.

4. Notes

1. X-Gluc is significantly less expensive if purchased in quantities of 10 g or more. The staining solution can be made in advance and stored in aliquots, wrapped in aluminum foil, at –20°C. Ferricyanide and ferrocyanide are included in the solution to increase the specificity of staining. These chemicals speed the oxidative dimerization of the product of the GUS enzymatic reaction (a soluble monomer) into a nondiffusible dimer, however, they are also inhibitory to the GUS enzyme, resulting in decreased sensitivity. A concentration of 2 m*M* works well for most purposes, but the concentration can be increased or decreased as needed. Chloramphenicol is included to inhibit enzymatic activity from contaminating bacteria during the incubation. To minimize costs, the GUS stain solution can be reused once, except in those cases where strong GUS activity has turned the stain blue, then it should be discarded.
2. Singer and Burke (Chapter 15 this text) have described a modified high-throughput TAIL-PCR protocol that utilizes pools of AD primers in a single reaction and only two rounds of PCR.
3. Occasionally, a product will not be visible in the secondary products of a successful TAIL-PCR. If a clear product is present in the tertiary reaction, you may wish to determine its sequence. It is also common for more than one product to be amplified. In this case, you may see multiple bands in the secondary and tertiary TAIL-PCR products.
4. Many universities and a number of commercial facilities offer DNA sequencing services. Contact the facility for template and primer concentration requirements.

   Cloning the TAIL-PCR product prior to sequencing is an alternative to direct sequencing and often results in higher quality sequence. This step is recommended when multiple products are amplified in the TAIL-PCR. After purification, estimate the concentration of the PCR product by running 2 μL of the product on an agarose gel. Clone the PCR products using the pGEM®-T Easy vector system II (Promega) following the manufacturer's instructions. Isolate plasmids from the bacterial cultures using a QIAprep plasmid miniprep kit. Verify the presence of an insert by restriction digestion with *Eco*RI. Prior to sequencing, the authenticity and orientation of TAIL-PCR products can be determined by PCR, by amplification with the appropriate *Ds* primer (either Ds3-4 or Ds5-4) in combination with either the M13 forward or M13 reverse primer. This step will discard any PCR products that did not result from direct amplification of flanking *Ds* sequence.
5. We have observed cases in which *GUS* expression is observed from a *DsG* element that is inserted so that *GUS* is in the opposite orientation to a gene in the region, and no genes in the correct orientation can be found. This may be explained by the presence of cryptic promoter sequences in the end of the *Ds* element ***(35)***, which effectively allow the *DsG* to act as an enhancer trap. Other examples of expressed promoter- or gene-trap insertions in regions where no detectable transcribed gene is present have also been reported ***(36–39)***. Reporter gene activation in these cases has been interpreted as being due to the activation of a cryptic promoter in the genome.

6. We do not know the maximum distance over which plant enhancer elements are able to act. Because the *Arabidopsis* genome is densely packed, with approx 1 gene every 5 kb ***(40)***, it is generally assumed that regulatory elements do not act over long distances.
7. It may be possible to detect transcripts using *in situ* RT-PCR, which provides significantly more sensitivity than conventional *in situ* hybridization ***(41,42)***.
8. There are several appropriate vectors available for this purpose, and the choice of vector is often based on the availability of suitable restriction enzyme sites. See Hellens et al. ***(43)*** for a good description of available binary vectors.
9. It may be wise to first determine whether the mutation is linked to the transposable element. Linkage can be examined in a population that segregates for both the mutant phenotype and for the transposable element. To simplify the analysis, verify that there is a single *Ds* element present in the population before proceeding. You may need to outcross the homozygous transposant line to wild-type plants and self the resulting progeny to obtain such a population. Score at least 100 individual families (each derived from a single plant) for the presence of the mutation. For a simple recessive mutation, you can determine whether each family is homozygous for the wild-type allele, heterozygous, or homozygous for the mutant allele. This should be done in the absence of selection for the transposable element. Independently score each family for the presence of the *Ds* element by selection on kanamycin. Examine the data to determine if the *Ds* element co-segregates with the mutation. If the mutation is due to the *Ds* insertion (or tightly linked), there should be a complete correlation between the presence of the mutation and presence of the *Ds* element. If you identify any families where the mutant allele is present in the absence of the *Ds* element (100% kanamycin-sensitive), then the mutation cannot be due to the *Ds* insertion. Moreover, you should not observe families that contain the *Ds* element but not the mutant allele. This analysis will allow you to determine whether the mutation is linked to the *Ds* element, but does not definitively prove that the *Ds* element is the cause of the mutation.

   If the mutant phenotype can be reverted by excision of the *Ds* element in either germinal or somatic cells, that is a good indication that the *Ds* insertion is the cause of the phenotype. Before initiating this experiment, you should verify that both ends of the *Ds* element are intact by amplification and sequencing. Transposition can result in disruption of one or both *Ds* terminal inverted repeats, and if this happens, the *Ds* element will have lost its ability to transpose ***(44)***. If the *Ds* ends are intact, then remobilization should be possible. Remobilization of the *Ds* element should be initiated by crossing the homozygous transposant to a line carrying *Ac* transposase to generate F_1 seed. Grow F_1 plants, allow them to self-pollinate, and harvest F_2 seed. Plant F_2 seed and identify those plants that are homozygous for the mutation. Among these homozygotes, identify those that contain the transposase gene using PCR amplification with primers AC1 5'-TAAAGCCGAGGAGTGGAAGA-3' and AC2 5'-TCCCCTCCACCATGATA AAA-3', which are specific to the *Ac* element and do not amplify the *Ds* element.

If the *Ds* insertion and the *Ac* transposase are unlinked, 75% of *Ds* homozygous plants should carry the transposase. Depending on the phenotype, it may be possible to identify revertant sectors in the somatic tissue of these F_2 plants. In many cases however, this will not be possible. To obtain germinal revertants, allow homozygous mutants containing the transposase to self-pollinate and examine the F_3 generation for the presence of wild-type individuals. At this stage, extreme care must be taken to avoid contamination from wild-type seed stocks, as these plants would appear to be phenotypic revertants. A *Ds* insertion almost always results in the generation of an 8-bp target-site duplication at the site of insertion, and *Ds* excision typically leaves behind a footprint, which results from imprecise excision ***(45)***. Therefore, you can confirm that the phenotypically wild-type plants observed in the F_3 generation are the result of a *Ds* excision event and are not due to contaminating wild-type seed by examining the insertion site for the presence of a footprint. Amplify and sequence genomic DNA flanking the *Ds* insertion site in phenotypically wild-type F_3 individuals. Detection of a footprint indicates that the wild-type phenotype results from excision of the *Ds* element and is not due to contamination from other wild-type seed stocks and that the mutation is due to the *Ds* insertion.

Depending on the location of the *Ds* insertion, an exact excision may be required to restore gene function. Within the coding region, excisions that leave the reading frame intact (insertions or deletions of 3 bp or multiples thereof) will be required. Some regions of a protein may not tolerate the insertion or deletion of even a single amino acid. If the original *Ds* insertion was located in a noncoding region of the gene, then a wider variety of excision events are likely to result in phenotypic reversion.

Acknowledgments

The authors thank members of the Springer laboratory for advice and discussions, and Sonia Zarate for comments on the manuscript. Research related to gene traps has been supported by grants from the National Science Foundation (IBN-9875371) and the Southwest Consortium (99-N02) to P.S.S.

References

1. Casadaban, M. J. and Cohen, S. N. (1979) Lactose genes fused to exogenous promoters in one step using a Mu-*lac* bacteriophage: *in vivo* probe for transcriptional control sequences. *Proc. Natl. Acad. Sci. USA* **76,** 4530–4533.
2. Bellen, H. J. (1999) Ten years of enhancer detection: lessons from the fly. *Plant Cell* **11,** 2271–2281.
3. Springer, P. S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell* **12,** 1007–1020.
4. Stanford, W. L., Cohn, J. B., and Cordes, S. P. (2001) Gene trap mutagenesis: past, present and beyond. *Nat. Rev. Genet.* **2,** 756–768.
5. Chin, H. G., Choe, M. S., Lee, S. H., et al. (1999) Molecular analysis of rice plants

harboring an *Ac/Ds* transposable element-mediated gene trapping system. *Plant J.* **19,** 615–623.
6. Martirani, L., Stiller, J., Mirabella, R., et al. (1999) T-DNA tagging of nodulation- and root-related genes in *Lotus japonicus*: expression patterns and potential for promoter trapping and insertional mutagenesis. *Mol. Plant Microbe Interact.* **12,** 275–284.
7. Nishiyama, T., Hiwatashi, Y., Sakakibara, K., Kato, M., and Hasebe, M. (2000) Tagged mutagenesis and gene trap in the moss, *Physcomitrella patens* by shuttle mutagenesis. *DNA Res.* **7,** 9–17.
8. Jeon, J. S., Lee, S., Jung, K. H., et al. (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J.* **22,** 561–570.
9. Martienssen, R. A. and Springer, P. S. (2000) Enhancer and gene trap transposon mutagenesis in *Arabidopsis*, in (http://www.arabidopsis.org/info/springer.html).
10. Weigel, D. and Glazebrook, J. (2002) *Arabidopsis: A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.
11. Tsugeki, R., Kochieva, E. Z., and Fedoroff, N. V. (1996) A transposon insertion in the *Arabidopsis SSR16* gene causes an embryo-defective lethal mutation. *Plant J.* **10,** 479–489.
12. Liu, Y.-G., Mitsukawa, N., Oosumi, T., and Whittier, R. F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* **8,** 457–463.
13. Church, G. M. and Gilbert, W. (1984) Genomic sequencing. *Proc. Natl. Acad. Sci. USA* **81,** 1991–1995.
14. Shure, M., Wessler, S., and Fedoroff, N. (1983) Molecular identification and isolation of the *Waxy* locus in maize. *Cell* **35,** 225–233.
15. Martienssen, R. A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl. Acad. Sci. USA* **95,** 2021–2026.
16. Sambrook, J. and Russell, D. W. (2001) *Molecular Cloning: A Laboratory Manual*. CSH Laboratory Press, Cold Spring Harbor, NY.
17. Ride, J. P., Davies, E. M., Franklin, F. C., and Marshall, D. F. (1999) Analysis of *Arabidopsis* genome sequence reveals a large new gene family in plants. *Plant Mol. Biol.* **39,** 927–932.
18. Cock, J. M. and McCormick, S. (2001) A large family of genes that share homology with *CLAVATA3*. *Plant Physiol.* **126,** 939–942.
19. MacIntosh, G. C., Wilkerson, C., and Green, P. J. (2001) Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.* **127,** 765–776.
20. Llave, C., Kasschau, K. D., Rector, M. A., and Carrington, J. C. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14,** 1605–1619.
21. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002) MicroRNAs in plants. *Genes Dev.* **16,** 1616–1626.
22. Seki, M., Narusaka, M., Kamiya, A., et al. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296,** 141–145.

23. Jefferson, R. A., Kavanagh, T. A., and Bevan, M. W. (1987) GUS fusions: β-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* **6,** 3901–3907.
24. Sieburth, L. E. and Meyerowitz, E. M. (1997) Molecular dissection of the *AGAMOUS* control region shows that *cis* elements for spatial regulation are located intragenically. *Plant Cell* **9,** 355–365.
25. Brand, U., Grünewald, M., Hobe, M. and Simon, R. (2002) Regulation of *CLV3* expression by two homeobox genes in *Arabidopsis*. *Plant Physiol.* **129,** 565–575.
26. Lohmann, J. U., Hong, R. L., Hobe, M., et al. (2001) A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. *Cell* **105,** 793–803.
27. Xiang, C., Han, P., Lutziger, I., Wang, K., and Oliver, D. J. (1999) A mini binary vector series for plant transformation. *Plant Mol. Biol.* **40,** 711–717.
28. Nagel, R., Elliott, A., Masel, A., Birch, R. G., and Manners, J. M. (1990) Electroporation of binary Ti plasmid vector into *Agrobacterium tumefaciens* and *Agrobacterium rhizogenes*. *FEMS Microbiol. Lett.* **67,** 325–328.
29. Clough, S. J. and Bent, A. J. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743.
30. Baxter-Burrell, A., Yang, Z., Springer, P. S., and Bailey-Serres, J. (2002) RopGAP4-dependent Rop GTPase rheostat control of *Arabidopsis* oxygen deprivation tolerance. *Science* **296,** 2026–2028.
31. Feldmann, K. A. (1991) T-DNA insertion mutagenesis in *Arabidopsis*: mutational spectrum. *Plant J.* **1,** 71–82.
32. Bancroft, I., Jones, J. D. G., and Dean, C. (1993) Heterologous transposon tagging of the *DRL1* locus in *Arabidopsis*. *Plant Cell* **5,** 631–638.
33. Sundaresan, V., Springer, P., Volpe, T., et al. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9,** 1797–1810.
34. Benfey, P. N., Ren, L., and Chua, N. H. (1989) The CaMV 35S enhancer contains at least two domains which can confer different developmental and tissue-specific expression patterns. *EMBO J.* **8,** 2195–2202.
35. Cocherel, S., Perez, P., Degroote, F., Genestier, S., and Picard, G. (1996) A promoter identified in the 3' end of the *Ac* transposon can be activated by *cis*-acting elements in transgenic *Arabidopsis* lines. *Plant Mol. Biol.* **30,** 539–551.
36. Fobert, P. R., Labbé, H., Cosmopoulos, J., et al. (1994) T-DNA tagging of a seed coat-specific cryptic promoter in tobacco. *Plant J.* **6,** 567–577.
37. Foster, E., Hattori, J., Labbé, H., et al. (1999) A tobacco cryptic constitutive promoter, *tCUP*, revealed by T-DNA tagging. *Plant Mol. Biol.* **41,** 45–55.
38. Plesch, G., Kamann, E., and Mueller-Roeber, B. (2000) Cloning of regulatory sequences mediating guard-cell-specific gene expression. *Gene* **249,** 83–89.
39. Mollier, P., Hoffmann, B., Orsel, M., and Pelletier, G. (2000) Tagging of a cryptic promoter that confers root-specific *gus* expression in *Arabidopsis thaliana*. *Plant Cell Reports* **19,** 1076–1083.
40. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815.

41. Ruiz-Medrano, R., Xoconostle-Cazares, B., and Lucas, W. J. (1999) Phloem long-distance transport of *CmNACP* mRNA: implications for supracellular regulation in plants. *Development* **126,** 4405–4419.
42. Koltai, H. and Bird, D. M. (2000) High throughput cellular localization of specific plant mRNAs by liquid-phase in situ reverse transcription-polymerase chain reaction of tissue sections. *Plant Physiol.* **123,** 1203–1212.
43. Hellens, R., Mullineaux, P., and Klee, H. (2000) A guide to *Agrobacterium* binary Ti vectors. *Trends Plant Sci.* **5,** 446–451.
44. Coupland, G., Plum, C., Chatterjee, S., Post, A., and Starlinger, P. (1989) Sequences near the termini are required for transposition of the maize transposon *Ac* in transgenic tobacco plants. *Proc. Natl. Acad. Sci. USA* **86,** 9385–9388.
45. Baker, B., Schell, J., Lörz, H., and Fedoroff, N. (1986) Transposition of the maize controlling element "Activator" in tobacco. *Proc. Natl. Acad. Sci. USA* **83,** 4844–4848.

15

High-Throughput TAIL-PCR as a Tool to Identify DNA Flanking Insertions

Tatjana Singer and Ellen Burke

Summary

Thermal asymmetric interlaced polymerase chain reaction (TAIL-PCR) is a fast and efficient method to amplify unknown sequences adjacent to known insertion sites in *Arabidopsis.* Nested, insertion-specific primers are used together with arbitrary degenerate primers (AD primers), which are designed to differ in their annealing temperatures. Alternating cycles of high and low annealing temperature yield specific products bordered by an insertion-specific primer on one side and an AD primer on the other. Further specifity is obtained through subsequent rounds of TAIL-PCR, using nested insertion-specific primers. The increasing availability of whole genome sequences renders TAIL-PCR an attractive tool to easily identify insertion sites in large genome tagging populations through the direct sequencing of TAIL-PCR products. For large-scale functional genomics approaches, it is desirable to obtain flanking sequences for each individual in the population in a fast and cost-effective manner. In this chapter, we describe a TAIL-PCR method amenable for high-throughput production (HT-TAIL-PCR) in *Arabidopsis* ***(1)***. Based on this protocol, HT-TAIL-PCR may be easily adapted for other organisms.

Key Words

TAIL-PCR, HT-TAIL-PCR, T-DNA, transposon, *Arabidopsis*, high-throughput, reverse genetics, tagging population, knock-out

1. Introduction

With the advancement of whole genome sequencing projects in plants, insertional mutagenesis has become an increasingly attractive tool to establish gene function through loss-of-function alleles (*see* Chapters 3, 13, and 17 in this text). In this reverse genetics approach, either endogenous or foreign DNA sequences serve as "tags" that, once inserted into a gene, disrupt its function.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

DNA sequences, such as transposable elements or T-DNA, the portion of the tumor-inducing (Ti) plasmid from *Agrobacterium* that is transferred into plant cells, have become the most widely employed insertional mutagens in plants ***(2)***. Because the sequence of the mutagen is known, it may be used as a bait to obtain sequences flanking the insertion. If the entire genome sequence is available, the exact location of the insertion is easily determined through homology searches (i.e., Basic Local Alignment Search Tool [BLAST] ***[3])*** against a whole genome database.

In large-scale functional genomics approaches, which aim to saturate the genome with insertions, large tagging populations are generated (*see* Chapter 19 in this text), and thousands of insertion sites have to be determined. The success of such a large-scale project depends critically on the ability to amplify sequences flanking insertion sites in a cost-effective and high-throughput manner. Polymerase chain reaction (PCR) products have to be of sufficient length and quality to serve directly as templates for sequencing and, thus, to reliably identify the insertion site through a BLAST search. The recovered flanking sequences are finally stored in a database that allows the researcher to search for knock-outs in genes of interest *in silico*.

Thermal asymmetric interlaced PCR (TAIL-PCR) ***(4)*** has been proven an efficient and sensitive method to amplify unknown sequences adjacent to tagged sites ***(5–10)***. Compared to other methods, like inverse PCR or adapter-ligation PCR, TAIL-PCR has a number of advantages that facilitate and expedite the procedure of retrieving sequences flanking insertion sites. For example, neither DNA manipulations, such as restriction cutting or adapter ligation, are required prior to TAIL-PCR nor are laborious screenings of PCR products necessary afterwards. Moreover, TAIL-PCR yields products of sufficient length and purity, which are ready for direct sequencing.

With the development of high-throughput technology, it has become feasible to analyze large collections of genome-tagged plants in a short time. In this chapter, we describe a high-throughput TAIL-PCR (HT-TAIL-PCR) protocol developed for an *Arabidopsis* knock-out population in which T-DNA was used as a mutagen. However, this protocol may easily be adapted to other species or different insertion tags, such as transposable elements. Where applicable, we will refer to these modifications in the Notes section.

1.1. Principles of TAIL-PCR

The key feature of the TAIL-PCR strategy is the use of two primer-sets that differ in length and have different melting temperatures (thermal asymmetry) ***(4)***. One set of primers consists of long nested primers complementary to the known insertion sequence that have high melting temperatures. Those primers are designed to read outwards from the known sequence into the unknown

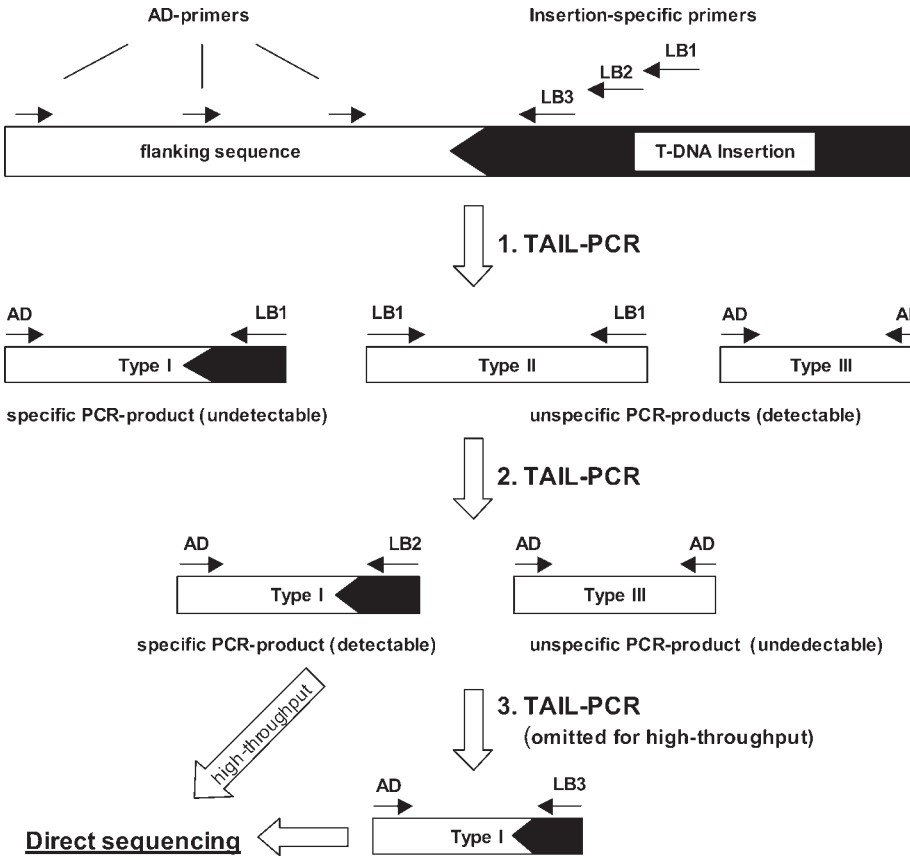


Fig. 1. Schematic representation of primer binding sites in TAIL-PCR and TAIL-PCR products. One side of the T-DNA insertion and the T-DNA-specific insertion primers (LB1, LB2, LB3) are shown examplary for any insertional mutagen used for tagging purposes. For clarity only, one side of the T-DNA insertion and the LB primers are depicted. Similar products would be expected with T-DNA RB primers. For further explanations, see text.

flanking sequence (*see* **Fig. 1**). The other primer set consists of arbitrary degenerate primers (AD primers), which are shorter and, therefore, anneal at lower temperatures. Depending on their level of degeneracy, AD primers are able to hybridize at random to many sites in the genome (*see* **Fig. 1**). High annealing temperatures, therefore, favor hybridizing of the long insertion-specific primers, whereas at low annealing temperatures, both primer sets hybridize with similar efficiency. In the TAIL-PCR protocol, interlaced cycles of high and low stringency take advantage of the thermal asymmetry of the primers in order to amplify the preferred target product. Further specificity is

obtained through subsequent rounds of amplification with nested insertion-specific primers. To ensure that sufficient random priming throughout the genome occurs, AD primer concentration exceeds the insertion-specific primer concentration, thus increasing the chance of having an AD primer binding close to the insertion site.

Three types of products are expected to occur in TAIL-PCR: the preferred specific target products, referred to as type I products, are primed by the insertion-specific primer on one side and a nonspecific AD primer on the other side. Type II products are nonspecific products, primed on both sides by the insertion-specific primer. Another class of nonspecific products, termed type III, are primed on both sides by nonspecific AD primers.

In order to increase the amount of specific template, the primary round in the TAIL-PCR protocol starts with 5 cycles with high annealing temperature (i.e., high stringency) (*see* **Fig. 2A**). This allows binding of the insertion-specific primer and, through linear PCR amplification, production of specific single-stranded product. In the second step, a cycle with low annealing temperature (i.e., low stringency) allows hybridization and extension of nonspecific AD primers. Because of the lower temperature, mismatch pairing is allowed, thus generating AD primer-specific target sites for the next round.

In the subsequent TAIL-cycling rounds, the preferred target molecule (type I) is amplified together with nonspecific type II and type III products (*see* **Fig. 1**). TAIL cycling consists of 15 super cycles, in which each super cycle consists of two high stringency cycles and one low stringency cycle (*see* **Fig. 2A**).

During the high stringency cycles, the insertion-specific primer binds preferably to its target sequences, and the resulting product is linearly amplified (thermal asymmetry). During the reduced stringency cycle, the lower annealing temperature allows also the AD primers to bind (thermal symmetry), and the high stringency products are converted into double-stranded form. Thus, the number of specific target molecules for the next round increases logarithmically. The final PCR usually consists of a mixture of high amounts of nonspecific type II product (primed on both sides by insertion-specific primers), moderate amounts of specific type I target product, and low amounts of type III nonspecific products (primed on both sides by AD primers). After the first TAIL-PCR round, multiple bands, mostly representing nonspecific type II products, are usually visible on an agarose gel, whereas the abundance of the specific type I product is too low to be detected.

In order to increase the yield of the specific type I product and to decrease the amount of contaminating nonspecific products, a second round of TAIL-PCR is performed using a nested insertion-specific primer (*see* **Fig. 2B**). The nested primer is designed to hybridize to the known insertion sequence internally of the first primer (*see* **Fig. 1**). Because type II nonspecific products were

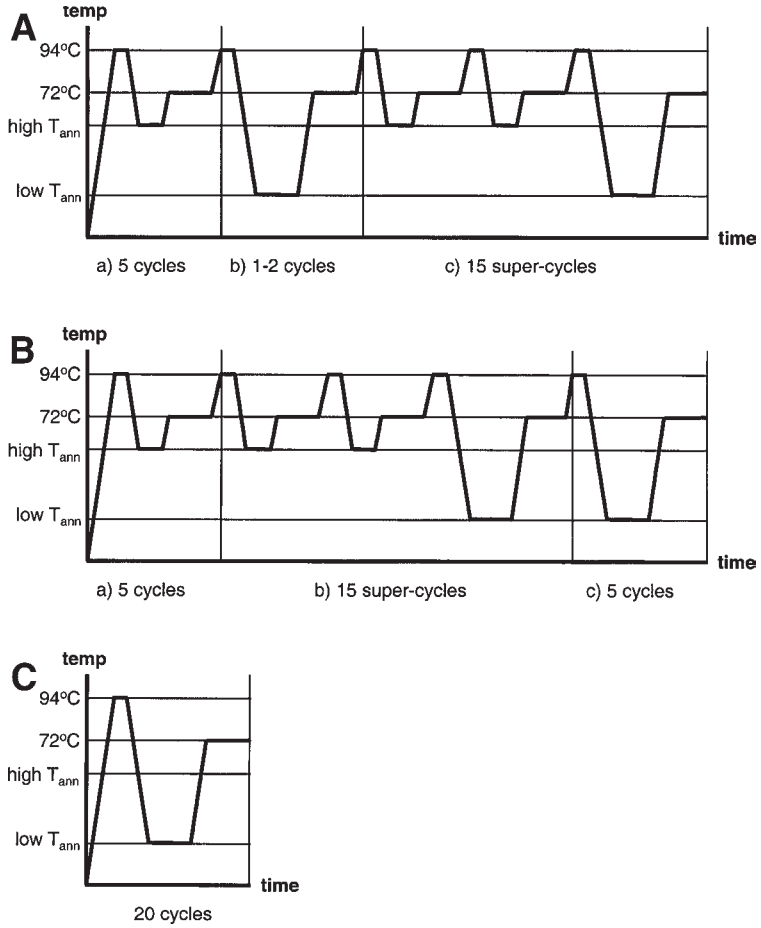


Fig. 2. Schematic representation of temperature profiles for HT-TAIL-PCR cycling. **(A) Primary HT-TAIL-PCR.** (a) 5 cycles with annealing temperature allow binding of insertion-specific primers. Specific single stranded product is produced. (b) 1-2 cycles of low annealing temperature allows binding of non specific AD-primers. AD primer-specific target sites are generated. (c) TAIL-cycling. two high-stringency cycles alternate with one low-stringency cycle for 15 super-cycles. Specific target molecules are produced together with nonspecific products. **(B) Secondary HT-TAIL-PCR.** (a) 5 cycles with high annealing temperature allow binding of insertion-specific primers. Specific single stranded product is pruduced. (b) 15 super-cycles of TAIL-cycling ensure production of preferred target molecule over nonspecific products. (c) 5 cycles at low annealing temperature increase amount of product for direct sequencing in HT-TAIL-PCR (*see* **Note 28**). **(C) Tertiary TAIL-PCR.** 20 cycles at low annealing temperature to increase amplification of specific products. For HT-TAIL-PCR the third round of TAIL-PCR is omitted (*see* **Note 28**). For detailed explanations see text. high T_{ann}: high stringency annealing temperature. low T_{ann}: low stringency annealing temperature.

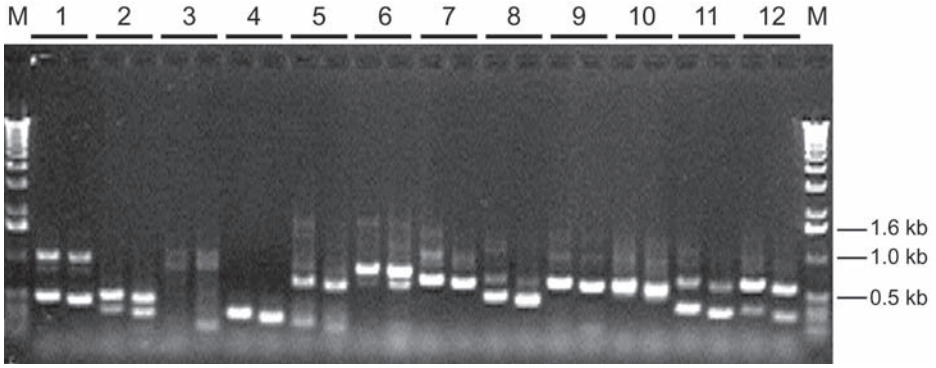


Fig. 3. One percent agarose gel stained with ethidium bromide. Twelve samples of secondary and tertiary TAIL-PCR products are loaded next to each other. Note the slight decrease in size of tertiary TAIL-PCR products. Multiple bands may be caused through multiple T-DNA insertions or originated through nested AD primer binding. M, marker (1-kb ladder).

generated through mispriming of insertion-specific primers in the first round, those products fail to reamplify with the nested primers in the second round.

Dilution of the primary TAIL-PCR followed by several super-cycles of TAIL-cycling (two high stringency cycles interlaced with one low stringency cycle) in the second round ensure that amplification of specific type I products is increased over contaminating type III products (*see* **Fig. 2B**). After the second round of TAIL-PCR, the yield of specific target products is sufficiently high to become visible on agarose gels. The amount of type III products should be extremely low and no longer detectable.

In the standard TAIL-PCR protocol, the secondary round of TAIL-amplification is followed by a third round using another set of nested insertion-specific primers (*see* **Fig. 2C**). Dilution of the secondary PCR products prior to the reaction and employing only a limited number of high stringency cycles (20) ensures that all undesired products fail to amplify. Through visual inspection on agarose gels of secondary and tertiary PCR products loaded next to each other, specific products are corroborated through a decrease in product size according to the position of the nested primers (*see* **Fig. 3**).

Multiple bands are often observed after secondary or tertiary TAIL-PCRs (*see* **Fig. 3**). Those bands are indicative of either multiple insertion sites of the mutagenic DNA into the genome or represent one insertion site primed with nested AD primers. Because the latter products represent the same flanking sequence of an insertion site, but only differ in length, sequencing of those products with the insertion-specific primer as a sequencing primer usually results in clean sequencing profiles.

1.2. HT-TAIL-PCR for T-DNA Insertion Lines

In order to decrease the number of PCRs and to increase efficiency, the original TAIL-PCR protocol was modified to be amenable for high-throughput purposes ***(1)***. Prior to HT-TAIL-PCR, genomic DNA from individual insertion lines is prepared to serve as template in the subsequent PCRs. In the protocol described below, we will describe a method for plant DNA extraction in 96-well format. In the standard TAIL-PCR protocol, usually six reactions with one insertion-specific primer and each of the six AD primers are performed for three rounds of PCR to maximize the likelihood of generating a specific product. Therefore, 18 PCRs have to be performed to amplify products adjacent to any T-DNA insertion. If reactions are done for both flanking sequences, the number of PCRs doubles. For a knock-out collection of approx 100,000 *Arabidopsis* plants, this would amount to 3,600,000 total PCRs. Analysis of the secondary and tertiary PCR products on agarose gels (24 reactions) would require 2,400,000 lanes. For larger mutant populations of plants, these numbers would increase substantially.

In the HT-TAIL-PCR protocol, the standard TAIL-PCR procedure has been modified, so that sequences flanking T-DNA insertions are amplified in only two reactions per plant line. The major modifications of the procedure are that *(i)* a pool of four AD primers is used together with the insertion-specific primer; *(ii)* only insertion-specific primers for one side of the T-DNA (left border) are used; and *(iii)* the third round of PCR is omitted. Employing this method, 2 to 3 products are produced on average, ranging in size from 100–1500 bp (*see* **Fig. 3**). Visual inspection of PCR products on agarose gels is only performed for optimization purposes, while the procedure is established, or occasionally during the process of high-throughput analysis in order to evaluate the quality and reliability of the PCRs. After the second round of HT-TAIL-PCR, the whole reaction mixture is purified from primers and excessive nucleotides through hydrolytic enzyme digestion and sequenced directly. With an automated BLAST search of the retrieved TAIL sequences against the whole genome sequence, insertion sites may be precisely mapped. Sequence reads and information about the insertion site are stored in a database that may be queried by the user for lines with insertions in genes of interest.

2. Materials

2.1. Tissue Collection and Disruption

1. Centrifuge for microtiter plates: Models 4-15C or 4K 15C (Qiagen).
2. Retsch MM300 Mixer Mill (Qiagen).
3. Mixer Mill adapter set 2× 96 (Qiagen).
4. Tungsten Carbide Beads 3 mm (Qiagen), (other beads are available from Retsch).

5. 96-Well tissue collection plates, containing 1.2-mL collection tubes with 8-strip microtube caps (Qiagen).
6. Microseal 'P' Sealing Pad (MJ Research), cut to size to fit Qiagen tissue collection plates.

2.2. Plant DNA Preparation

1. Insulated laboratory pans (ice buckets) (41 × 33 cm) (Fisher Scientific).
2. Hydra-96 Microdispenser (290-μL needle vol) (Robbins Scientific, Apogent Discoveries).
3. Four Qfill2 dispensers (Genetix).
4. Four Qfill2 dispenser caps (Genetix).
5. Eight 500-mL glass bottles.
6. Glass bead-filled dry bath (Fisher Scientific).
7. 12-Multichannel Impact Pipetor (15–850 μL) (Matrix Technologies, Apogent Discoveries).
8. Talltips™ extended length filter tips (1250 μL) (Matrix Technologies, Apogent Discoveries).
9. 100-mL Reagent reservoirs (sterile) (Matrix Technologies, Apogent Discoveries).
10. Sterile polypropylene U-bottom 96–well plates (Matrix Technologies, Apogent Discoveries).
11. Tape-pad (Qiagen).
12. Absolute ethanol (200 proof).
13. DNeasy 96 Plant Kit (6) (Qiagen).
14. 1.2-mL Collection tubes.
15. Square well blocks (2 mL).
    All kit reagents, plates, tubes, and caps may be ordered separately from Qiagen.
16. Hoefer DyNA Quant 200 Fluorometer (Amersham Pharmacia Biotech).
17. Hoechst Dye 33258 (Amersham Pharmacia Biotech).
18. Calf thymus DNA (Sigma).

2.3. Oligonucleotides

1. Oligonucleotides are ordered at 200 nmol scale, desalted (Invitrogen or various suppliers).
2. TE: 10 m*M* Tris-HCl, pH 8.0, 1 m*M* ethylenediamine tetraacetic acid (EDTA).
3. AD primer: *see* **Table 1**.
4. T-DNA-specific primers: *see* **Table 2**.
5. Dissociation element-specific primers: *see* **Table 3.**

2.4. HT-TAIL-PCR

1. Model 9700 Thermal cycler (holds 2× 384-well plates) (Applied Biosystems).
2. Hydra-384 Microdispenser (100-μL needle vol) (Robbins Scientific, Apogent Discoveries).
3. 384-Well plates for PCR (Applied Biosystems).

Table 1
AD Primer Sequences

Primer name	Primer sequence (5'–3')	Length	Degeneracy	Average T_m	Average GC	Reference
AD1[a]	NGTCGASWGANAWGAA	16 bp	128-fold	46.6°C	43.8%	AD2 in ***(5)***
AD2[a]	TGWGNAGSANCASAGA	16 bp	128-fold	49.2°C	50%	AD1 in ***(4)***
AD3[a]	AGWGNAGWANCAWAGG	16 bp	128-fold	46.6°C	43.8%	AD2 in ***(4)***
AD4	STTGNTASTNCTNTGC	16 bp	256-fold	47.9°C	46.9%	AD5 in ***(9)***
AD5	NTCGASTWTSGWGTT	15 bp	64-fold	43.7°C	43.3%	AD1 in ***(5)***
AD6[a]	WGTGNAGWANCANAGA	16 bp	256-fold	45.3°C	40.6%	AD3 in ***(5)***
			total averages:	46.6°C	44.7%	

[a]Only those AD primers are used in the protocol and pooled.
W = A or T, S = G or C, N = A or T or G or C.

Table 2
T-DNA Primer Sequence

Left T-DNA border[a]	Primer sequence (5'–3')	Length	T_m	GC	Reference
LB1[b]	GCCTTTTCAGAAATGGATAAATAGCCTTGCTTCC	34 bp	67.0°C	41.2%	***(6)***
LB2[b]	GCTTCCTATTATATCTTCCCAAATTACCAATACA	34 bp	63.5°C	32.4%	***(6)***
LB3[b]	TAGCATCTGAATTTCATAACCAATCTCGATACAC	34 bp	64.7°C	35.3%	***(6)***
Right T-DNA border[c]					
QRB1	CAAACTAGGATAAATTATCGCGCGCGGTGTC	31 bp	68.2°C	48.4%	***(6)***
QRB2	GGTGTCATCTATGTTACTAGATCGGGAATTGA	32 bp	65.6°C	40.6%	***(6)***
QRB3	CGCCATGGCATATGCTAGCATGCATAATTC	30 bp	66.8°C	46.7%	***(6)***
Right T-DNA border[d]					
RB1	ATTAGGCACCCCAGGCTTTACACTTTATG	29 bp	65.3°C	44.8%	***(6)***
RB2	GTATGTTGTGTGGAATTGTGAGCGGATAAC	30 bp	65.4°C	43.3%	***(6)***
RB3	TAACAATTTCACACAGGAAACAGCTATAC	29 bp	61.0°C	34.5%	***(6)***

[a]LB primer specific for T-DNA vectors pCSA110 and pDAP101.
[b]The described TAIL-PCR protocol has been optimized for those primers.
[c]RB primers specific for T-DNA vector pCSA110.
[d]RB primer specific for T-DNA vector pDAP101. T-DNA transformation vectors pCSA110 and pDAP101 are described in ref. ***6***.

Table 3
Ds Element Primer Sequences

Ds 3'-end	Primer sequence (5'–3')	bp	T_m	GC	Reference
Ds3'-1	GGTTCCCGTCCGATTTCGACT	21 bp	61.8°C	57.1%	***(9)***
Ds3'-2	CGATTACCGTATTTATCCCGTTC	23 bp	58.9°C	43.5%	***(9)***
Ds3'-3	TCGTTTCCGTCCCGCAAGT	19 bp	58.8°C	57.9%	***(9)***
Alternative primer set:					
Ds3'-1	CGATTACCGTATTTATCCCGTTCG	24 bp	61.0°C	45.8%	***(19)***
Ds3'-2	CCGGTATATCCCGTTTTCG	19 bp	56.7°C	52.6%	***(19)***
Ds3'-3	GAAAATGAAAACGGTAGAGGT	21 bp	54.0°C	38.1%	***(19)***
Ds 5'-end					
Ds5'-1	ACGGTCGGGAAACTAGCTCTAC	22 bp	62.1°C	54.5%	***(9)***
Ds5'-2	TCCGTTCCGTTTTCGTTTTTTAC	23 bp	62.4°C	52.2%	***(9)***
Ds5'-3	CGGTCGGTACGGGATTTTCC	20 bp	61.4°C	60.0%	***(9)***
Alternative primer set:					
Ds5'-1	CCGTTTACCGTTTTGTATATCCCG	24 bp	61.0°C	45.8%	***(19)***
Ds5'-2	CGTTCCGTTTTCGTTTTTTACC	22 bp	56.5°C	40.9%	***(19)***
Ds5'-3	CGGTCGGTACGGGATTTTCC	20 bp	61.4°C	60%	***(19)***

4. 96-Pin replicators with OmniTray plate copier (Nalgene Nunc International).
5. 100-mL Reagent reservoirs (sterile).
6. 12-Channel Impact Pipetor (2–125 µL range) (Matrix Technologies, Apogent Discoveries).
7. Sterile filter tips (200 µL) (Matrix Technologies, Apogent Discoveries).
8. MicroAmp clear adhesive films (Applied Biosystems).
9. Sealing roller (MJ Research).
10. Microseal 'P' sealing pads (MJ Research), cut to size to fit Model 348 PCR-block (Applied Biosystems).
11. 15-mL Falcon® tubes (Fisher Scientific).
12. PCR-grade water (Invitrogen).
13. Platinum-*Taq* DNA polymerase (Invitrogen).
14. 10× PCR-buffer (Invitrogen).
15. $MgCl_2$ (50 m*M*) (Invitrogen).
16. dNTPs (100 m*M*) (e.g., Applied Biosystems, New England Biolabs, Stratagene, Roche Molecular Biochemicals).
17. 2% Bleach solution (prepare fresh).
18. 96-Well plates (Fisher Scientific).
19. Single-Channel Impact Repeating Pipetor (5–250 µL range) (Matrix Technologies, Apogent Discoveries).

20. 12-Channel Micropipet (0.2–10 μL range) (Fisher Scientific).
21. Sterile filter tips (250 μL) (Matrix Technologies, Apogent Discoveries).
22. Sterile filter tips (10 μL) (Fisher Scientific).

2.5. Agarose Gel Electrophoresis

1. TAE buffer: 242 g Tris base, 57.1 mL acetic acid, 100 mL 0.5 *M* EDTA, distilled water (dH_2O) to 1 L, pH 8.5.
2. 10× DNA gel loading buffer: bromphenol blue-xylene xyanol (for 100 mL): 250 mg bromphenol blue (Sigma), 250 mg xylene xyanol (Sigma), 33 mL 150 m*M* Tris-HCl, pH 7.6, 60 mL glycerol, dH_2O to 100 mL.
3. 5× DNA gel loading buffer: orange G (for 100 mL): 50 mL glycerol, 10 mL 0.5 *M* EDTA, 10 mL 2% orange G (Sigma), 1 mL 10% sodium dodecyl sulfate (SDS), 29 mL dH_2O.
4. Agarose (SeaKem®).
5. Agarose gel apparatus suited for multichannel-pipet loading, e.g., Econo-Sub gel chamber with six 27-well combs (2 mm) (C.B.S. Scientific).

2.6. Purification of PCR Products

1. ExoSAP-IT™ (USB) (contains exonuclease I and shrimp alkaline phosphatase [SAP]) or buy individual enzymes (Amersham Pharmacia Biotech). Store at –20°C.

2.7. Sequencing

1. Model 9700 Thermal cycler (Applied Biosystems).
2. Model 3700 Sequencer (capillary) (Applied Biosystems).
3. 384-Well PCR plates (Applied Biosystems).
4. Sequencing primer: LB3 primer.
5. BigDye™ Terminator v2.0 Cycle Sequencing Ready Reaction (contains AmpliTaq® DNA polymerase, FS) (Applied Biosystems).
6. 5× Sequencing-buffer: 2 m*M* $MgCl_2$, 80 m*M* Tris-HCl, pH 8.0.
7. Sephadex®-G50 (Sigma).
8. 384 Filter plates (Whatman).

3. Methods

3.1. Tissue Collection and Disruption

1. Collect 30–50 mg of leaf tissue (typically approx 2 to 3 medium sized *Arabidopsis* leaves) or 4 to 5 inflorescences and add to well of 1.2 mL 96-well Qiagen tissue collection plate. Repeat for all 96-wells (*see* **Note 1**).
2. Lyophilize tissue overnight and store at –80°C until ready for preparation.
3. Two tissue collection plates are processed at the same time. Place one tungsten carbide bead into each tube (*see* **Note 2**). Secure eight-strip microtube caps over each tube in a column. Place a Microseal sealing pad (cut to size to fit on 96-tube tissue collection rack) over the sealed tubes and put lid over the pad and tubes (*see* **Note 3**).

4. Pour enough liquid nitrogen into a styrofoam container, such that three-quarters of the plate is immersed and freeze plate for several seconds (*see* **Note 4**).
5. Place frozen collection plates between mixer mill adapter plates and secure safely in mixer mill. Disruption is carried out for 1.5 min at a frequency of 20–30 s. Remove one plate and make sure that the tissue has been disrupted into a fine powder and that all the caps are still placed securely. If the tissue has not been disrupted completely, freeze plates again and repeat disruption by placing the racks in opposite orientation in the mixer mill (*see* **Note 5**). Disrupted tissue is stored at –80°C. Work quickly during disruption procedure, do not allow tissue to thaw.

3.2. Plant DNA Preparation

Genomic DNA is extracted with the DNeasy 96 Plant kit. Four times 96-well plates are processed at the same time (*see* **Note 6**). All centrifugation steps are carried out at room temperature.

1. Preheat a glass bead-filled dry bath to 65°C (*see* **Note 7**).
2. In a 500-mL glass bottle mix 48 mL buffer AP1, 120 µL reagent DX, and 120 µL RNase A (100 mg/mL) per 96 samples. Put Qfill2 dispenser cap on bottle and place it into a glass bead-filled dry bath and incubate at 65°C. Prepare at least 15% more buffer than needed to account for loss.
3. Fill three glass bottles each with buffers AP2, AP3/E, and AW. Add required amount of ethanol to buffer AW. Place Qfill2 dispenser caps on bottles. Prepare four bottles filled with dH_2O.
4. Set up four Qfill2 dispensers in a row and program each Qfill dispenser according to the vol that has to be dispensed for each buffer: AP1 400 µL, AP2 130 µL, AP3/E 600 µL, AW 800 µL. Connect buffer bottles to Qfill2 machines. Purge buffers through Qfill2 dispensers in order to fill the system with liquid (*see* **Note 8**).
5. Place a 96-well rack containing the collection tubes with the pulverized leaf samples on first Qfill2 platform and dispense 400 µL of preheated AP1 buffer into each well. Recap the tubes with fresh eight-strip caps, place lid on box, and shake vigorously for several seconds (*see* **Note 9**).
6. Remove lid from box, fold paper towel, and place on top of the caps. Put lid back on and place box in 65°C water bath so that it is about half covered but not immersed. Place a round weight over the box and incubate for 15 min. After incubation, remove plate from the water bath and let cool for 5 min (*see* **Note 10**).
7. Quick-spin for 10 s at 2500*g* to force condensation droplets down.
8. Remove and discard caps. Place box with collection tubes on second Qfill2 platform and add 130 µL AP2 buffer to each well, close with new eight-strip caps. Cover with lid. Shake box up and down for 15 s. Quick-spin rack of collection tubes for 10 s, 2500*g*, and incubate for 10–15 min at –20°C (*see* **Note 11**).
9. Shake box for a few seconds and centrifuge for 5 min at 5600*g* (*see* **Note 12**).

10. While plates are spinning, place DNeasy 96 plate over a 96-well plate. Place plates on third Qfill2 dispenser platform and add 600 µL of AP3/E buffer to the DNeasy 96 plate (*see* **Note 13**).
11. After centrifugation, use the 12-channel mulitpipet and the extended filter tips to transfer 400 µL of supernatant into the buffer-filled DNeasy 96 plate. Secure a tape-pad over the plate and mix gently inverting the plate 5 to 6×. Place DNeasy 96 plate on square well block and centrifuge for 4 min at 5600*g* (*see* **Note 14**). Discard flow-through and rinse out square well block for reuse.
12. Remove tape pad and place DNeasy 96 plate over empty 96-well plate. Position plates on fourth Qfill2 dispenser platform. Dispense 800 µL AW buffer into each well (*see* **Note 15**). Place DNeasy 96 plate on square well block. Centrifuge for 4 min at 5600*g*. Discard flow-through (*see* **Note 16**).
13. Place the DNeasy 96 plate on top of a fresh sterile 96-well plate and incubate in a dry incubator at 70°C for 10 min. After incubation, make sure all the ethanol is evaporated (no moisture should be visible on the sides of the wells) (*see* **Note 17**).
14. Prepare Hydra-96 Microdispenser for dispensing AE buffer. Add buffer to appropriate reagent reservoir and take up 100 µL. Program Hydra-96 for dispensing of 2× 50 µL (*see* **Note 18**).
15. Place DNeasy 96 plate over of 96-well plate. Place plates on working platform of the Hydra-96 and dispense 50 µL. Wait for 1 min and centrifuge plates for 2 min at 5600*g*.
16. Repeat elution process with 50 µL AE buffer.
17. Remove DNeasy 96 plate and place tape pad over the 96-well plate. Label the plate for identification and place at –80°C for long-term storage. DNA may be stored at 4°C for several days.
18. DNA quality, yield, and concentration are examined on 1% TAE agarose gels (*see* **Note 19**). Before gel electrophoresis, spin plates for 20 s at 2500*g* to force down condensation droplets. Load 10 µL of DNA sample with 1 µL 10× bromphenol blue–xylene xyanol gel loading buffer in each well (*see* **Note 20**). Use appropriate length marker standard. Electrophoresis should be performed at 6 V/cm in 1× TAE buffer. Take photograph. Reseal plates with fresh tape pads.

3.3. Oligonucleotides

1. Order AD primers and insertion-specific primers. For comments on primer design, *see* **Note 21**. If necessary, design insertion-specific oligonucleotide primers (*see* **Note 22**) tailored to your T-DNA (*see* **Note 23**) or for your specific insertional mutagen (*see* **Note 24**). Different AD primers may have to be designed for plant species other than *Arabidopsis* in order to account for differences in GC content of the respective plant genome (*see* **Note 25**).
2. Resuspend lyophilized oligonucleotides in TE or dH_2O. Flick gently, let dissolve at room temperature for at least 2 h or at 4°C overnight, spin down briefly.
3. Prepare 100 µ*M* stock solution of LB primers (*see* **Note 26**).
4. Prepare 200 µ*M* stock solution for AD primers. Store primer stock solutions at –20°C.

Table 4
AD-Pool Primer Concentrations

AD Primer	4× AD-pool working solution (sufficient for two 384-well plates)	Concentration in 4× AD-pool
AD1	300 µL	12 µM
AD2	300 µL	12 µM
AD3	300 µL	12 µM
AD6	400 µL	16 µM
H_2O	3700 µL	
Total vol	5 mL	

5. Dilute stock solutions for insertion-specific primers 1:10 (e.g., three nested LB primers) to obtain 10 µ*M* working solutions. Keep on ice while preparing TAIL-PCR master mixture. Diluted primer working solutions are stored at 4°C.
6. Prepare 4× working solution of pooled AD primers (AD-pool) from 200 µM AD primer stock solutions (*see* **Table 4** and **Note 27**). Keep on ice while preparing TAIL-PCR master mixture. AD-pool working mixture is stored at 4°C.

3.4. HT-TAIL-PCR

3.4.1. General Comments and Preparations for HT-TAIL-PCR

1. When TAIL-PCR will be established in plants other than *Arabidopsis* or with different primer sets, it is advisable to optimize the procedure in 96-well format before scaling up to a high-throughput 384-well format. For comments on optimizing HT-TAIL-PCR, *see* **Note 28**.
2. In order to facilitate calculations for primer concentrations and master mixtures create Excel® spreadsheets to automatically calculate all concentrations and vol accordingly.
3. Cut microseal sealing pads to fit lid size of a thermal cycler.
4. Prepare fresh 2% bleach solution.
5. Prepare sufficient AD-pool working solution (*see* **Table 4**) for number of plates processed.
6. To prepare 100 m*M* dNTP stock solution, mix equal amounts of individual 100 m*M* dNTP stock solutions (dATP, dTTP, dCTP, dGTP) together. The final concentration of each dNTP in the stock mixture is then 25 m*M*. Dilute the 100 m*M* dNTP stock mixture 1:10 to obtain a 10 m*M* dNTP working solution. dNTP stock mixture (100 m*M*) and 10 m*M* dNTP working solution are stored at –20°C.
7. Aliquot reagents (dNTPs, 10× PCR buffer, $MgCl_2$, primer) in 15-mL Falcon tubes and keep on ice. Before use, thaw all reagents on ice, vortex mix, and spin down briefly. Master mixtures are prepared in 100-mL sterile reagent reservoirs and mixed thoroughly by gently moving the tray back and forth. Keep master mixtures on ice.

Table 5
Single Reaction for Primary TAIL-PCR

1× Reaction vol	Reagents/stock solutions	Final concentration
4.7 μL	dH_2O	
1.0 μL	10× PCR buffer	1×
0.2 μL	10 m*M* dNTP working solution	0.2 m*M*
0.3 μL	50 m*M* $MgCl_2$	1.5 m*M*
0.2 μL	10 μ*M* LB1 primer	0.2 μ*M*
2.5 μL	4× AD-pool	1× (3–4 μ*M*)
0.1 μL	5 U/μL Platinum *Taq*	0.5 U
1 μL DNA	Template	
10 μL	Total vol	

8. Use PCR-grade water for all PCRs.
9. Quick spins are done for 10 s at 2500*g*.
10. Always prepare 15–20% more solution for master mixtures to account for loss in handling.

3.4.2. Primary HT-TAIL-PCR

1. Program Model 9700 thermal cyclers (*see* **Note 29**) for primary HT-TAIL-PCR (*see* **Subheading 3.5.1.** and **Note 30**; **Fig. 2A**).
2. Prepare master mixture for primary TAIL-PCR (*see* **Table 5**).
3. Select 96-well plates containing the genomic DNA to be processed (4× 96-well DNA plates per 384-well plate). Thaw plates on ice, spin down briefly.
4. Label a set of four 96-well DNA plates A, B, C, and D that will be replicated into one 384-well PCR plate.
5. Label 384-well plates. Write on one side of the plate "Experiment_ID TAIL1, date" and on the other side note which quadrant of the 384-well plate corresponds to which 96-well DNA plate (e.g., A: 96-well DNA plate no. #; B: 96-well DNA plate no. #; C: 96-well DNA plate no. #; D: 96-well DNA plate no. #). Mark the A1 corner on the plate.
6. Prepare four wash trays: fill one reservoir with the freshly prepared 2% bleach solution (half full) and the remaining three with dH_2O (nearly full). Label the reservoirs accordingly (*see* **Note 31**).
7. Distribute 9 μL of master mixture into each well of the 384-well plates using a 12-channel multipipet. Keep the plates on ice until ready to use.
8. To transfer DNA from 96-well DNA plates to 384-well plates, place the 384-well plate in the plate copier (white plastic frame with quadrants A, B, C, and D designated). Insert 96-pin replicator into the A-designated 96-well DNA plate. Move the replicator to the 384-well plate guided to the A quadrant by the two outside pins. Let the pins rest in the wells for several seconds, gently move around, and then remove. Place the 96-pin replicator into the 2% bleach reservoir.

9. Repeat the replication process for quadrant B with the plate designated B and a new 96-pin replicator. When done with quadrant B, move the pin-replicator used for plate A out of the bleach reservoir and into the first water reservoir. After it stands for approx 1 min, make quick final rinses in the next two water reservoirs. Shake excess water off and then set the replicator down on its side on a paper towel. Air-dry thoroughly before using again. Put second pin-replicator into bleach reservoir. Repeat the replication-cleaning process for the remaining C and D plates (*see* **Note 32**).
10. Place a 384-well MicroAmp adhesive film over the 384-well reaction plate and seal with a roller.
11. While processing the other 384-well plates in the set, place the completed reaction plates on ice or at 4°C.
12. When all the plates in the set are ready for thermal cycling, spin the plates briefly.
13. Place two reaction plates in each Model 9700 thermal cycler, such that the A1 position is at the upper left. Place microseal sealing pads (cut to size) over the top of the plates and close the lid. Start program TAIL1. The program is completed in approx 4 h.

3.4.3. Dilution of Primary HT-TAIL-PCR

Prior to the secondary HT-TAIL-PCR, the primary HT-TAIL-PCR is diluted 1:100, and 1 µL of that dilution is used as template for the second reaction. Plates (2 × 4, 384-well) are processed at the same time.

1. Prepare eight new 384-well plates for a 1:100 dilution of the primary TAIL-PCRs. In order obtain a 1:100 dilution, 0.2 µL of the primary PCR is mixed with 19.8 µL of dH_2O using the Hydra-384 Microdispenser (*see* **Note 33**).
2. Label 384-well plates for the 1:100 dilutions. Write on one side of the plate: "TAIL1 @ 1:100, date" and on the other side note which quadrant of the 384-well plate corresponds to which 96-well DNA plate. Mark the A1 corner on the plate.
3. Prepare automatic pipeting machine: turn on Hydra-384 Microdispenser and empty the water from the syringes. Briefly raise the water reservoir to rinse water drops from the syringe tips. Program Hydra-384 accordingly (*see* **Note 34**).
4. Remove the water reservoir. Place a sterile wash tray labeled "PCR water" on the stage. Fill the tray with PCR-grade water.
5. Go to File 1. Scroll to D 0.0 and press fill. The syringes are programmed to fill to 100 µL.
6. Remove the wash tray and place the first 384-dilution plate on the stage. Press dispense (red button on the right). The program is set to dispense 19.8 µL. Repeat for three remaining 384 dilution plates. Empty any water remaining in syringes into the wash tray and repeat the process for the other four plates.
7. Remove the primary TAIL-PCR plates from the thermal cyclers. Centrifuge briefly. Carefully remove the sealing tape from the plates.
8. In order to dispense 0.2 µL of the primary TAIL-PCR into the dilution plates, set

Table 6
Single Reaction for Secondary TAIL-PCR

1× Reaction vol	Reagents/stock solutions	Final concentration
5.7 µL	dH_2O	
1.0 µL	10× PCR buffer	1×
0.2 µL	10 m*M* dNTP working solution	0.2 m*M*
0.3 µL	50 m*M* $MgCl_2$	1.5 m*M*
0.2 µL	10 µ*M* LB2 primer	0.2 µ*M*
1.5 µL	4× AD-pool	0.6× (1.8 – 2.4 µ*M*)
0.1 µL	5 U/µL Platinum *Taq*	0.5 U
1 µL 1:100 diluted first TAIL	Template	
10 µL	Total vol	

airgap-function of Hydra-384 to 4 µL (*see* **Note 35**). Go to File 3 and scroll down to A 0.0. Press the red aspirate button on the right to fill with 4 µL of air.

9. Press the left, red reset button and go to File 2. Place a primary TAIL-PCR plate on the Hydra stage and scroll to A 4.0. Press the red aspirate button to fill syringes with 0.2 µL. The Hydra should now read A 4.2. Remove the plate and place the corresponding dilution plate on the stage. Press empty to dispense the 0.2 µL into the dilution plate.
10. Remove the dilution plate. Place a tape pad on the plate and seal with a roller. Set plate aside on ice or at 4°C until the other plates are done. Plates with diluted DNA may be stored at 4°C for a few days.
11. Fill wash tray half with freshly prepared 2% bleach solution. Clean the Hydra syringes by placing the reservoir on the stage and press the wash button. When the wash cycle is complete, remove the bleach tray. Place a fresh wash tray filled with dH_2O on the stage. Press wash. When the wash cycle is complete, remove the tray and spill out the water. Fill tray with fresh water and repeat the wash cycle. The Hydra is now ready for the next 384 primary TAIL-PCR plate (*see* **Note 36**).
12. Place the next 384 primary TAIL-PCR plate on the stage and repeat dilution **steps 4–10**. Repeat for all the other plates. When all the plates are done, centrifuge plates briefly. Set dilution plates on ice or keep at 4°C until ready to use.

3.4.4. Secondary HT-TAIL-PCR

1. Program Model 9700 thermal cyclers for secondary HT-TAIL-PCR (*see* **Fig. 2B** and **Subheading 3.5.2.**).
2. Prepare master mixture for secondary TAIL-PCR (*see* **Table 6**).
3. Label eight new 384-well plates for secondary TAIL-PCR. On one side, write "Experiment_ID TAIL2, date" and on the other side note which quadrant of the

384-well plate corresponds to which 96-well DNA plate. Mark the A1 corner on the plate.
4. Distribute 9 μL of master mixture into each well of the 384-well plates using a 12-channel multipipet. Keep plates on ice until ready to use.
5. In order to dispense 1 μL of the primary TAIL-PCR dilution into the master mixture, set the airgap-function of Hydra-384 to 4 μL: go to File 3 and scroll down to A 0.0. Press the red aspirate button to fill with 4 μL of air (*see* **Note 37**).
6. Press the reset button and go to File 4. Place a primary TAIL-PCR dilution plate on the Hydra stage and scroll to A 4.0. Press the aspirate button to fill syringes with 1.0 μL. The Hydra should now read A 5.0. Remove the dilution plate and place the corresponding secondary TAIL-reaction plate on the stage. Press empty to dispense 1 μL into the wells.
7. Remove secondary TAIL-reaction plate from Hydra-384 stage. Place MicroAmp adhesive film over the plate and seal firmly with a roller. Keep plate on ice until other plates are done.
8. Wash Hydra-384 syringes with 2% bleach solution and water as described in **Subheading 3.4.3.**, **step 11**.
9. Repeat the procedure for the remaining plates.
10. When finished with the Hydra-384 for the day, close it down by placing a wash reservoir on the stage (filled with dH_2O). Go to File 1 and run a wash cycle. When the syringes fill with water, turn the machine off.
11. Once all the secondary TAIL-PCR plates are ready, centrifuge briefly and place in Model 9700 thermal cyclers, such that the A1 position is at the upper left. Place microseal sealing pads (cut to size) over top of the plates and close the lids. Start program TAIL2. The program should run approx 4.5 h.

3.4.5. Tertiary HT-TAIL-PCR

For high-throughput production of TAIL-PCR products for sequencing the third round of TAIL-PCR (**Fig. 2C**) is omitted (*see* **Note 28**).

1. Program Model 9700 thermal cyclers for tertiary HT-TAIL-PCR (*see* **Subheading 3.5.3.**).
2. Prepare Master Mixture for tertiary TAIL-PCR (*see* **Table 7**).
3. Prepare a 1:50 dilution of secondary TAIL-PCR and label plates accordingly. Keep dilution plates on ice.
4. Prepare fresh 96-well plates for a tertiary round of TAIL-PCR and label them accordingly.
5. Use a single-channel repeating pipet to dispense 19 μL of master mixture into each well of the tertiary 96-well plates. Add 1 μL of diluted secondary TAIL-PCR to the master mixture using a 12-channel micropipet.
6. Seal plates with adhesive film, spin quickly, and place in a thermal cycler.
7. Run program TAIL3. The program is completed in approx 2–2.5 h.

Table 7
Single Reaction for Tertiary TAIL-PCR

1× Reaction vol	Reagents/stock solutions	Final concentration
10.4 µL	dH_2O	
2.0 µL	10× PCR buffer	1×
0.4 µL	10 m*M* dNTP working solution	0.2 m*M*
0.6 µL	50 m*M* $MgCl_2$	1.5 m*M*
0.4 µL	10 µ*M* LB3 primer	0.2 µ*M*
5.0 µL	4× AD-pool	1× (3–4 µ*M*)
0.2 µL	5 U/µL Platinum *Taq*	1 U
1 µL 1:50 diluted third TAIL	Template	
20 µL	Total vol	

3.5. PCR Programs

3.5.1. First Round HT-TAIL-PCR (TAIL 1)

Use heated lid.

Program Model 9700 thermal cycler for first round of HT-TAIL-PCR (*see* **Note 29**) (**Fig. 2A**). (If MJ Research DNA machines are used for thermal cycling, program them according to **Note 38**).

1. 94°C for 3 min (*see* **Note 39**).
2. 94°C for 30 s.
3. 62°C for 1 min.
4. 72°C for 2:30 min.
5. Five cycles of **steps 3–5** (*see* **Note 40**).
6. 94°C for 30 s.
7. 25°C for 3 min (50% ramp).
8. 72°C for 2:30 s (32% ramp).
9. Two cycles of **steps 7–9** (*see* **Note 41**).
10. 94°C for 10 s.
11. 68°C for 1 min.
12. 72°C for 2:30 min.
13. 94°C for 10 s.
14. 68°C for 1 min.
15. 72°C for 2:30 min.
16. 94°C for 10 s.
17. 44°C for 1 min.
18. 72°C for 2:30 min.
19. 15 cycles of **steps 11–19** (*see* **Note 42**).

20. 72°C for 5 min.
21. 4°C hold.

3.5.2. Second Round HT-TAIL-PCR (TAIL 2)

Use heated lid.
Program thermal cyclers for second round of HT-TAIL-PCR (*see* **Fig. 2B**).

1. 94°C for 3 min.
2. 94°C for 10 s.
3. 64°C for 1 min.
4. 72°C for 2:30 min.
5. Five cycles of **steps 2–4** (*see* **Note 43**).
6. 94°C for 10 s.
7. 64°C for 1 min.
8. 72°C for 2:30 min.
9. 94°C for 10 s.
10. 64°C for 1 min.
11. 72°C for 2:30 min.
12. 94°C for 10 s.
13. 44°C for 1 min.
14. 72°C for 2:30 min.
15. 15 cycles of **steps 6–9** (*see* **Note 44**).
16. 94°C for 10 s.
17. 44°C for 1 min.
18. 72°C for 3 min.
19. Five cycles of **steps 16–18** (*see* **Note 45**).
20. 72°C for 5 min.
21. 4°C hold.

3.5.3. Third Round HT-TAIL-PCR (TAIL 3)

Use heated lid.
Program thermal cyclers for third round of TAIL-PCR (*see* **Fig. 2C**).

1. 94°C for 3 min.
2. 94°C for 10 s.
3. 44°C for 1 min.
4. 72°C for 2 min.
5. 20 cycles of **steps 2–4** (*see* **Note 46**).
6. 72°C for 5 min.
7. 4°C forever.

3.6. Agarose Gel Electrophoresis

Primary TAIL-PCRs usually are not analyzed on agarose gels, because no specific products are visible after the first round. In order to verify if specific

Table 8
ExoSAP Master Mixture

1× Reaction vol	Reagents	Final concentration/13-μL reaction
0.25 μL	Exonuclease I (10 U/μL)	2.5 U
0.25 μL	SAP (2 U/μL)	0.5 U
2.50 μL	dH_2O	
3.0 μL	Total vol	

products have been amplified, secondary and tertiary TAIL-PCRs are loaded side-by-side on agarose gels. Specific type I products are characterized through a decrease in size after the third round, corresponding to the basepair difference of the nested primers (*see* **Fig. 3** and **Notes 28** and **47**).

1. Transfer 10 μL of PCRs from secondary and tertiary TAIL-reaction plates (96-well format, 20 μL total vol), into fresh 96-well plates, add 2 μL of 5× orange G gel loading buffer. Orange G dye runs at the samplefront (<50 bp).
2. Prepare 1 to 2% agarose gel with 1× TAE buffer.
3. Use agarose gel system that allows 12-channel multipipet loading.
4. Load secondary and tertiary TAIL-PCR products in alternating wells, so that products may be examined next to each other. Load appropriate size markers.
5. Run gels at approx 100 V for approx 20–30 min and take photograph.

For troubleshooting, *see* **Note 48**.

3.7. Purification of TAIL-PCR Products

Secondary TAIL-PCRs are purified with ExoSAP treatment prior to sequencing (*see* **Note 49** and **Table 8**).

1. Prepare ExoSAP master mixture (*see* **Table 8**). Keep on ice.
2. Add 3 μL of ExoSAP master mixture directly to each 10-μL secondary TAIL-PCR in 384-well plate with a 12-channel multipipet. Seal plates with adhesive film and centrifuge plates briefly.
3. Program Model 9700 thermal cycler for ExoSAP program: 37°C for 20 min, 80°C for 15 min, and 4°C hold.
4. Place plates into thermal cycler and start ExoSAP program.
5. After program is finished, TAIL-PCRs are ready for sequencing.

3.8. Sequencing of TAIL-PCR Products

1. Sequence secondary HT-TAIL samples using standard sequencing procedures (*see* **Note 50**). Use LB3 primer or appropriate insertion-specific primer as the sequencing primer. Sequencing reactions are set up in 384-well plates. Prepare sequencing master mixture (*see* **Table 9**). Keep on ice.

Table 9
1× Sequencing Reaction

1× Reaction vol	Reagents
1.0 μL	BigDye Terminator V.2
0.5 μL	5× buffer
0.25 μL	Sequencing primer (LB3) 10 μ*M*
1.75 μL	dH_2O
1.5 μL	DNA (purified secondary TAIL-PCR reaction)
5 μL	Total vol

2. Program Model 9700 thermal cycler for cycle-sequencing program: 95°C for 15 s, 50°C for 5 s, and 60°C for 2 min, for 25 cycles, then 4°C hold. Sequencing products are purified with Sephadex G-50 columns.
3. Hydrate 70 g Sephadex G50 with 1 L dH_2O for at least 4 h at room temperature or overnight at 4°C.
4. Using an 8-channel multipipet, dispense 100 μL of hydrated Sephadex per well into 384-well filter plate and place filter plate on fresh 384-well collection plate.
5. Spin at 910*g* for 5 min. Discard flow-through contained in collection plate.
6. Pipet another 50 μL of hydrated Sephadex resin into each well. Spin for 5 min at 910*g* and discard flow-through contained in lower collection plate.
7. Add 5 μL dH_2O to 5 μL sequencing reaction in 384-well plates.
8. Place plate containing Sephadex resin on fresh 384-well plate. Take entire 10 μL vol of sequencing reaction and pipet on top of the Sephadex columns in the 384-well plates.
9. Spin at 910*g* for 5 min. Seven to eight microliters of clean sequencing products are usually recovered.
10. Analyze 7–8 μL of cleaned up sequencing reactions on ABI 3700 sequencers (Capillary).

4. Notes

1. Because genomic plant DNA is prepared in 96-well format using the DNeasy 96 Plant DNA Isolation kit, tissue collection is facilitated if plants are grown in 48-pot flats. Two flats of plants are then collected into one 96-well rack of tubes. Tracking of plants and seeds after harvesting is facilitated if individual plants are bar-coded.
2. We designed a bead dispenser that could be filled with beads and was constructed so as to dispense one bead for each of the 96 wells of the collection plate at the same time. Stainless steel beads may be used alternatively to tungsten carbide beads.
3. The sealing pad ensures that the lid of the tissue collection plate fits tightly onto the caps of the collection tubes during disruption in mixer mill.

4. Be careful handling liquid nitrogen. Wear cryoprotective gloves and safety goggles. Be careful that no liquid nitrogen enters the collection tubes. If tubes contain liquid nitrogen, caps may explode after tissue disruption, and tissue powder may cross-contaminate to other tubes.
5. In order to check for tissue disruption, take plates out of mixer mill and knock the plates several times gently on the benchtop in order to bring disrupted tissue to the bottom of the tubes.
6. Four 96-well plates may be processed by one person at a time (in approx 3 to 4 h) using one centrifuge for 2× 96-well plates by efficiently staggering the workflow. Therefore, eight plates (768 DNA samples) may be prepared per day and person.
7. A glass bead-filled dry bath is conveniently used to avoid water contamination of AP1 buffer or floating bottles. Alternatively, a regular water bath may be used.
8. Because manifolds of Qfill2 dispensers may become clogged, it is important to check if manifolds are properly discharging liquid while the buffer is purged through the system. After use, Qfill2 dispensers are cleaned with dH_2O. Connect bottles containing dH_2O to the machines by placing the Qfill2 dispenser caps from the buffer bottles on them. Flush dH_2O through system.
9. At this step, addition of AP1 buffer lyses cell membranes, and DNA is released. RNA is digested with RNAse.
10. Because strip caps easily pop open during incubation at 65°C, it is important to ensure that the plate is not fully immersed in water and that a weight is put on top of the lid to avoid dilution and cross-contamination of samples. The paper towel that is placed between the caps and the lid absorbs any moisture and prevents spilling of liquid to neighboring tubes if caps pop open.
11. In this step, proteins and polysaccharides are precipitated.
12. Proteins and polysaccarides are removed by centrifugation.
13. The 96-well plate serves as a support platform for the DNeasy 96 plate to place it on the Qfill2 platform. The DNeasy 96 plate is presoaked with binding buffer AP3/E to calibrate the column material and to insure binding of the DNA.
14. Through addition of the supernatant to binding buffer AP3/E, DNA is precipitated and binds to column resin. Supernatant should be colorless, but may be of green color.
15. The wash buffer AW removes contaminating salts and residual buffer.
16. For centrifugation, DNeasy 96 plate is put on square well block. Do not centrifuge 96-well plates with DNeasy 96 plates.
17. At this step, it is crucial that all the ethanol is evaporated, because it would inhibit downstream TAIL-PCRs.
18. Familiarize yourself with Hydra-96 microdispenser and program machine accordingly. Reagent reservoirs may be obtained from the manufacturer, or lids of tip boxes may be used alternatively. DNA is eluted from DNeasy columns with AE buffer.
19. It is recommended to control yield and quality of DNA while the technique is established in the laboratory. Once DNA extraction is routinely done in high-

throughput format, checking of all samples on agarose gels is not necessary. However, few individual DNA samples from prepared plates may be checked randomly in order to ensure consistent quality. In addition to gel electrophoresis, spectrophotometric quantitation may be performed at OD_{260} or using a Fluorometer and Hoechst-Dye 33258 ***(11)***. However, visual inspection of genomic DNA on agarose gels should be preferred, because sample degradation may be examined. The DNA yield obtained by the described method ranges from approx 8–30 ng/µL (average approx 15 ng/µL) for *Arabidopsis*.

20. It is recommended to use gel chambers suited for alternating multichannel-pipet loading.
21. TAIL-PCR is based on thermal asymmetry of insertion-specific primers and AD primers. Melting temperatures (T_m) of insertion-specific primers should be at least approx 10°C higher than those of the AD primers. Typically, the T_m of AD primers is approx 46°C, and the T_m of insertion-specific primers should lie between 58°–65°C. Primer T_ms are calculated after the formula: 69.3 + 0.41(%GC) – 650/L ***(12)***. General rules for primer design should be considered (avoid hairpin structures, formation of primer–dimers, and GC-rich 3' ends).
22. In general, the insertion-specific primers should be located near the ends of the known insertion sequence, close to the point were the junction with the genomic sequence is expected to occur. The primers are designed in nested sets of three, with their 3' end facing outwards towards the border (*see* **Fig. 1**). The nested primers used in the primary and secondary TAIL-PCR should not overlap considerably. Especially the third primer used in the tertiary PCR should be located 60–90 bp away from the second primer, so that specific products may be identified through their size difference when run side-by-side on an agarose gel.
23. If primers are designed for T-DNA insertion. The nested insertion-specific primers should be located approx 100–200 bp internally to the 25-bp T-DNA border repeat. It should be ensured that the 3' end of the tertiary primer is at least 90–100bp bp internal to the T-DNA border repeat. Often, the T-DNA is not cut precisely during transfer, resulting in truncated inserts ***(13,14)***.
24. If transposable elements are used as insertional mutagen, insertion-specific primers are designed such that they are located in the terminal repeats of the element. Those primers may be located closer to the element ends, since deletion of transposon ends, during integration is rare. Since DS elements are a widely used mutagen in *Arabidopsis* ***(8)***, we have included published insertion-specific primers for this element in **Subheading 3** (*see* **Table 3**).
25. GC content of AD primers used in *Arabidopsis* is approx 45% on average (*see* **Table 1**) (*Arabidopsis* average genome-wide GC content is 35%). For plants, which differ considerably in GC content from *Arabidopsis*, AD primer sequences should be adjusted accordingly, but care has to be taken that thermal asymmetry between AD primers and insertion-specific primers is maintained. For other plant genomes, constitution of AD-pools may have to be modified, or AD primers may be used individually to obtain optimal results.
26. A disadvantage of using T-DNA as a insertional mutagen is that it is prone to

insert into the host plant genome in tandem arrays or more complex insertions patterns, resulting in adjacent T-DNA borders ***(15–17)*** and, therefore, hampering the isolation of plant sequences flanking the T-DNA inserts. Most of the adjacent T-DNA arrays in our mutant collection (62%) appear to be connected over the right borders (RB), compared to 25% that are connected over left border (LB) sequences ***(1)***. To increase efficiency and process large sample numbers in a high-throughput fashion, we, therefore, only used insertion-specific primers binding to the left border of the T-DNA (i.e., LB primers) in this protocol (*see* **Table 10**).

27. Instead of performing different TAIL-PCRs with individual AD primers, it was found that, for *Arabidopsis*, pooling of four AD primers (AD1, AD2, AD3, AD6) is the ideal combination yielding the most specific products from various combinations tested ***(1)***. The AD primers are pooled such that their final concentration in the reaction mixture for the primary and secondary HT-TAIL-PCR is proportional to their level of degeneracy (*see* **Table 11**). AD primer concentration is reduced in the second amplification in the HT-TAIL-PCR protocol, because it yields the best results for the 384-well format and 10-μL reaction vol ***(1)***.

    If AD primers are used individually, their concentration should be the same as indicated for the primary PCR in all three rounds of TAIL-PCR (*see* **Table 11**). Besides *Arabidopsis,* the AD-pool used in this protocol has been shown to work also for rice and soybean (E. Burke, personal communication).

28. Optimizing of TAIL-PCR should be done in 96-well format and 20-μL reaction vol. Ten microliters of secondary and tertiary TAIL-PCR is then analyzed with agarose gel electrophoresis (*see* **Fig. 3**), while the remainder may be used for sequencing. If TAIL-PCRs are performed in 96-well format, genomic DNA should be diluted 1:10 in order to obtain a suitable concentration of template for the primary TAIL-PCR. For 20-μL reaction vol, the recommended amount of genomic DNA used as a template is 1–5 ng (1 μL of 1:10 dilution) for *Arabidopsis*. For plants with larger genomes, more template DNA has to be used (approx 15–20 ng for rice and soybean, approx 30 ng for maize per 20-μL reaction) or the equivalent of 4000–8000 haploid genomes. For HT-TAIL-PCR done in 384-well format and 10-μL reaction vol, we found that no dilution of genomic DNA is necessary when a 96-pin replicator is used for transferring DNA into PCR reactions.

    In 96-well format, primary TAIL-PCRs are diluted 1:50, and 1 μL is then used as a template for the second round of TAIL-PCR (final dilution 1:1000). The same dilution is made for the secondary TAIL-PCR before the third round of TAIL-PCR. The third round of TAIL-PCR may be performed while the technique is established in the laboratory in order to optimize the protocol. In order to validate the amplification of specific products, secondary and tertiary PCR products are loaded on agarose gels next to each other (*see* **Fig. 3**). Specific tertiary products are smaller in size compared to secondary products. Once TAIL-PCR has been optimized in 96-well format, it may be scaled up to 384-well high-throughput format. The major changes in the described HT-TAIL-PCR protocol, compared to 96-well format, are that genomic DNA is not diluted before the first round of TAIL-PCR and that the primer concentration of the AD-pool is reduced in the

Table 10
T-DNA Primer Concentrations

T-DNA primer	Stock solution	Concentration in working solution	Final concentration in first PCR reaction	Final concentration in second PCR reaction	Final concentration in third PCR reaction
LB1*, LB2*, LB3*	100 μ*M*	10 μ*M*	0.2 μ*M*	0.2 μ*M*	0.2 μ*M*
RB1, RB2, RB3	100 μ*M*	10 μ*M*	0.2 μ*M*	0.2 μ*M*	0.2 μ*M*
QRB1, QRB2, QRB3					

Only primers marked with a * are used in this protocol. The same primer concentrations are used for other insertion-specific primers.

Table 11
AD Primer Concentrations

AD primer	Degeneracy	Stock solution	AD-pool: concentration in 4× working solution	Final concentration in first PCR reaction	Final concentration in second PCR reaction
AD5	65-fold	200 μ*M*	8 μ*M*	2 μ*M*	1.2 μ*M* (2 μ*M*)[a]
AD1*, AD2*, AD3*	128-fold	200 μ*M*	12 μ*M*	3 μ*M*	1.8 μ*M* (3 μ*M*)[a]
AD4, AD6*	256-fold	200 μ*M*	16 μ*M*	4 μ*M*	2.4 μ*M* (4 μ*M*)[a]

Only primers marked with a * are used in this protocol.
[a]AD primer concentrations for TAIL-PCR in 96-well format.

second round. Also, the third round of TAIL-PCR is omitted, and no agarose gel electrophoresis is performed. In order to obtain enough product for sequencing, five additional low stringency cycles are included in the secondary HT-TAIL-PCR after the TAIL-cycling rounds.

29. Alternatively to the Applied Biosystems 9700 machines, MJ Research thermal cyclers may be used for TAIL-PCR. If MJ Research thermal cyclers are used, it is required to purchase a PCR license from Applied Biosystems to be used with TAIL or other PCR methods.
30. The turnover of how many HT-TAIL-PCRs are being performed per day depends on the number of PCR machines available. To achieve a high-throughput of TAIL-PCRs, at least two 384-well thermal cyclers should used (for first and second round of HT-TAIL-PCR), processing four 384-well plates (1536 samples) at a time.
31. Wash trays are purchased by the manufacturer, or tip box lids may be used alternatively.
32. At this step, it is important to dry 96-pin replicators well before reuse. A blow dryer may be used to expedite drying.
33. Familiarize yourself with the Hydra-Microdispenser before use. Program files according to the model used. The Aspirate mode may vary between older and newer Hydra models.
34. Program Hydra-384: File 1: fill 100 μL, dispense 19.8 μL; File 2: aspirate 0.2 μL of DNA (in this model, airgap has to be set before it can be accessed); File 3: set airgap to 4 μL.
35. In order to accurately dispense vol of ≤1 μL from the Hydra-384 Microdispenser, it is recommended to use the airgap function at 4 μL (A 4.0). The airgap pushes the liquid out.
36. It is crucial to clean the Hydra syringes after processing each plate in order to prevent cross-contamination.
37. Program Hrydra-384: File 4: fill 1 μL, dispense 1 μL; File 3: set airgap to 4 μL.
38. Program MJ Research DNA machines according to protocol. Set Control Method to Calculated and use heated lid. Program **Subheading 3.5., steps 6–9** as follows:
    6. 94°C for 30 s.
    7. ramp 0.4°C/s to 25°C.
    8. 25°C for 3 min.
    8. ramp at 0.3°C/s to 72°C.
    9. 72°C for 2 min (for 1 cycle).
39. *Taq* DNA polymerase is activated through denaturing of Platinum *Taq* Antibody.
40. Five high stringency cyles favor production of single-stranded product, primed by the insertion-specific primer (*see* **Fig. 2A**).
41. Two low stringency cycles facilitate AD primer annealing. First, the annealing temperature is gradually decreased until it reaches 25°C (50% ramp), and then again, gradually increased (32% ramp) until it reaches extension temperature (72°C) (*see* **Fig. 2A**).

    In this protocol, two cycles are used because exact ramping times cannot be programmed on Applied Biosystems Model 9700 thermal cyclers as on MJ Research thermal cyclers.

42. Fifteen super-cycles of two high stringency cycles interlaced with one low stringency cycle (TAIL-cycling) (*see* **Fig. 2A**).
43. Five high stringency cycles favor production of single-stranded product, primed by the nested insertion-specific primer (*see* **Fig. 2B**).
44. During the 15 super-cycles, the specific target molecules (type I) are produced preferably. Because the nested insertion-specific primer is used, nonspecific type II products are suppressed. Also, type III contaminating products (primed on both sides by AD primers) are produced at significantly lower levels (*see* **Fig. 2B**).
45. Five cycles with low annealing temperature in order to increase amount of product for sequencing. These cycles may be omitted during optimization (*see* **Fig. 2B**).
46. Twenty cycles at low stringency temperature are used for amplification of specific products with third nested insertion-specific primer (*see* **Fig. 2C**).
47. During PCR, smaller products are more efficiently amplified than larger products. Therefore, smaller molecular weight bands are generally brighter, because they are present in greater amounts than larger products**.**
48. If a smear of PCR products is observed in secondary or tertiary TAIL-PCRs, it could be due to several reasons: *(i)* The $MgCl_2$ concentration in the reaction mixture is not optimal; *(ii)* dNTPs may have been misaliquoted; *(iii)* too little genomic DNA; or *(iv)* imbalance of primer concentrations. If there is not enough DNA in the reaction mixture, a smear may result because there is not enough template available to allow amplification of specific products. If there is too much DNA in the TAIL-PCR, this may result in no product because too many potential AD primer binding sites are available and, therefore, AD primers are titrated out. In HT-TAIL-PCR, too little AD-pool may cause smearing due to nonspecific amplification. If too much AD-pool is used in the secondary round of HT-TAIL-PCR, the concentration of residual primers still present after the reaction is quite high, causing problems for subsequent sequencing reactions.
49. Excess nucleotides and primers are digested by incubating the TAIL-PCR with exonuclease I and SAP. Both enzymes are active in the buffer used for PCR, so no change in buffer is required. Exonuclese I degrades primers and single-stranded DNA and SAP removes remaining dNTPs from the PCR mixture, which would interfere with the labeling step in the sequencing process. Both enzymes are heat-inactivated at 80°C.
50. Because HT-TAIL-PCRs are often a mixture of different products, the resulting sequence-reads may consist of chimeric sequences representing the multiple products in the mixture. Shorter HT-TAIL-PCR products are present at higher molar concentrations in the mixture, producing higher intensity signal peaks at the beginning of the sequence read. Longer products are present in lower molar concentrations, producing lower intensitiy signal peaks towards the end of the sequence read. If the sequence reads are not filtered for lower quality sequence (phred) ***(18)***, chimeric sequence reads may be identified on electropherograms. BLAST searches of chimeric sequences, containing up to four different products, confirmed independent insertion sites in the *Arabidopsis* genome (ref. ***1*** and G. Presting, personal communication).

References

1. Sessions, A., Burke, E., Presting, G., et al. (2002) A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell* **14,** 2985–2994.
2. Parinov, S. and Sundaresan, V. (2000) Functional genomics in *Arabidopsis*: large-scale insertional mutagenesis complements the genome sequencing project. *Curr. Opin. Biotechnol.* **11,** 157–161.
3. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
4. Liu, Y. G. and Whittier, R. F. (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics* **25,** 674–681.
5. Liu, Y. G., Mitsukawa, N., Oosumi, T., and Whittier, R. F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* **8,** 457–463.
6. McElver, J., Tzafrir, I., Aux, G., et al. (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis thaliana*. *Genetics* **159,** 1751–1763**.**
7. Budziszewski, G. J., Lewis, S. P., Glover, L. W., et al. (2001) *Arabidopsis* genes essential for seedling viability: isolation of insertional mutants and molecular cloning. *Genetics* **159,** 1765–1778.
8. Parinov, S., Sevugan, M., Ye, D., Yang, W.-C., Kumaran, M., and Sundaresan, V. (1999) Analysis of flanking sequences from dissociation insertion lines. A database for reverse genetics in *Arabidopsis*. *Plant Cell* **11,** 2263–2270.
9. Tsugeki, R., Kochieva, E. Z., and Fedoroff, N. V. (1996) A transposon insertion in the *Arabidopsis* SSR16 gene causes an embryo-defective lethal mutation. *Plant J.* **10,** 479–489.
10. Tissier, A. F., Marillonnet, S., Klimyuk, V., et al. (1999) Multiple independent defective *Suppressor-mutator* transposon insertions in *Arabidopsis*: a tool for functional genomics. *Plant Cell* **11,** 1841–1852.
11. Labarca, C. and Paigen, K. (1980) A simple, rapid, and sensitive DNA assay procedure. *Anal. Biochem.* **102,** 344–352.
12. Mazars, G.-R., Moyret, C., Jeanteur, P., and Theillet, C.-G. (1991) Direct sequencing by thermal asymmetric PCR. *Nucleic Acids Res*. **19,** 4783.
13. Gheysen, G., Herman, L., Breyne, P., Gielen, J., Van Montagu, M., and Depicker, A. (1990) Cloning and sequence analysis of truncated T-DNA inserts from *Nicotiana tabacum*. *Gene* **94,** 155–163.
14. Nacry, P., Camilleri, C., Courtial, B., Caboche, M., and Bouchez, D. (1998) Major chromosomal rearrangements induced by T-DNA transformation in *Arabidopsis*. *Genetics* **149,** 641–650.
15. De Neve, M., De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1997) T-DNA integration patterns in co-transformed plant cells suggest that T-DNA repeats originate from cointegration of separate T-DNAs. *Plant J.* **11,** 15–29.
16. Krizkova, L. and Hrouda, M. (1998) Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions. *Plant J.* **16,** 673–680.
17. De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1999) The DNA

sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. *Plant J*. **20,** 295–304.
18. Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185.
19. Grossniklaus, U., Vielle-Calzada, J. P., Hoeppner, M. A., and Gagliano, W. B. (1998) Maternal control of embryogenesis by MEDEA, a polycomb group gene in *Arabidopsis*. *Science* **280,** 446–450.

16

Custom Knock-Outs with Hairpin RNA-Mediated Gene Silencing

Susan Varsha Wesley, Qing Liu, Anna Wielopolska, Geoff Ellacott, Neil Smith, Surinder Singh, and Chris Helliwell

Summary

Hairpin (hpRNA)-mediated gene silencing exploits a cellular mechanism that recognizes double-stranded RNA (dsRNA) and subjects it and its corresponding mRNA to a sequence-specific degradation. This phenomenon is known as posttranscriptional gene silencing (PTGS) in plants and RNA interference (RNAi) in animals. dsRNA, when introduced into plant cells through hpRNA constructs, results in severe reduction of the target mRNA—the silencing effect being stably inherited over many generations. While hpRNA constructs can be made using conventional plasmids, use of generic vectors such as pHANNIBAL makes it more convenient to silence a number of genes simultaneously. Vectors, such as pHELLSGATE, that are based on the Gateway® technology are suitable for high-throughput gene silencing. The specificity of dsRNA silencing, it's ability to simultaneously silence multiple genes combined with the availability of high-throughput silencing vectors enables the researcher to generate custom knock-out plants.

Key Words

gene silencing, RNA interference, knock-out, hpRNA, pHANNIBAL, pHELLSGATE, Gateway, dsRNA

1. Introduction

Double-stranded RNA (dsRNA) is perceived by plant cells as foreign and triggers the degradation of itself and homologous RNA within the cell. Two protein complexes, DICER and RISC, are now implicated in chopping the dsRNA into small RNAs (small interfering RNA or siRNA of approx 21 bases long) and using them as guides to recognize corresponding mRNA for

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

sequence-specific degradation. This process is called posttranscriptional gene silencing (PTGS), which is also termed RNA interference (RNAi) in animals ***(1)***. Besides being a fascinating mechanism, it can be exploited as a functional genomics tool. Already, it has been used to ascertain the function of several genes in *Drosophila* and *Caenorhabditis elegans* ***(2,3)***.

Gene silencing can be achieved by transformation of plants with constructs that express self-complementary (termed hairpin [hp]) RNA containing sequences homologous to the target genes. The DNA sequences encoding the self-complementary regions of hpRNA constructs form an inverted repeat ***(4)***. The inverted repeat can be stabilized in bacteria through separation of the self-complementary regions by a "spacer" region. When the spacer sequence encodes an intron, the efficiency of gene silencing is very high, with up to 100% of the transformants generated with a particular gene construct showing some degree of silencing ***(5)***. There are at least three ways in which hpRNA constructs can be made. The construct may be generated from standard binary plant transformation vectors in which the hairpin-encoding region is generated *de novo* for each gene. Alternatively, generic gene silencing vectors such as the pHANNIBAL and the pHELLSGATE series ***(6,7)*** can be used. They simply require the insertion of polymerase chain reaction (PCR) products, derived from the target gene, into the vectors by conventional cloning or by using the Gateway® directed recombination system.

The features of intron interrupted hpRNA (ihpRNA)-mediated gene silencing make it particularly attractive in the production of knock-out plants for functional genomics applications. As hpRNA targets specific genes, each gene in the group under study can be targeted with a hpRNA construct. The hpRNA construct is genetically dominant, and therefore, phenotypes can be screened in primary transformed plants without the need to produce homozygous lines. Most plant transformation systems give rise to a number of transformation events that are propagated as separate transgenic lines. Thus, if a phenotype is replicated among the population of plants generated using a particular hpRNA transgene, it is highly likely that the phenotype is due to silencing of the target gene rather than caused by a mutation introduced by the transformation procedure. The differing degrees of silencing obtained in the lines produced from one transformation event may allow survival of weakly silenced lines for genes for which a complete loss of function would be lethal. The sequence specificity of gene silencing allows the use of unique sequences to target specific genes and the potential to use conserved sequences to target multigene families. This enables researchers to custom make knock-out plants to suit their requirements.

In this chapter, we describe three different ways of assembling hpRNA constructs using conventional plasmids, pH/KANNIBAL vectors, and pHELLSGATE

Table 1
Hairpin RNA Silencing

Gene (reference)	Species	Prom	Intron	Target	Stem (nt)	% Silenced prim. transf.	Construct type
PPO ***(6)***	Tobacco	35S	Pdk	ORF	572	70	ihp
GUS ***(6)***	Tobacco	35S	n/a	ORF	800	48	hp
PVY-Nia ***(6)***	Tobacco	35S	Pdk	ORF	730	58/96	hp/ihp
EIN2 ***(6)***	*Arabidopsis*	35S	Pdk	ORF	600	65	ihp
FLC1 ***(6)***	*Arabidopsis*	35S	Pdk	ORF	650	100	ihp
FLC1 ***(6)***	*Arabidopsis*	35S	Pdk	ORF	400	100	ihp
CHS ***(6)***	*Arabidopsis*	35S	Pdk	ORF	741	91	ihp
Δ12 ***(6)***	*Arabidopsis*	Napin	Δ12a	3'UTR	120	69/100	hp/ihp
AG ***(12)***	*Arabidopsis*	35S	n/a	ORF	554	99	hp
CLV 3 ***(12)***	*Arabidopsis*	35S	n/a	ORF	288	88	hp
AP 1 ***(12)***	*Arabidopsis*	35S	n/a	ORF	409	96	hp
PAN ***(12)***	*Arabidopsis*	35S	n/a	ORF	369	87	hp
CBL ***(13)***	*Arabidopsis*	35S	n/a	ORF	1146	91	hp
PDS ***(7)***	*Arabidopsis*	35S	Pdk	ORF	300	100	ihp
PhyB	*Arabidopsis*	35S	Pdk	3'UTR	300	70	ihp
Δ12 ***(11)***	Cotton	Lectin	n/a	ORF	853	58	hp
Δ12 ***(6)***	Cotton	Δ12c	Δ12c	5'UTR	98	100	ihp
Δ9 ***(11)***	Cotton	Lectin	n/a	ORF	514	57	hp
BYDV-Pol ***(14)***	Barley	Ubi	n/a	ORF	1600	36	hp
GUS ***(6)***	Rice	Ubi	n/a	ORF	560	85	hp

PPO, polyphenol oxidase; GUS, β-glucuronidase; PVY-NIa, potato virus Y NIa; ORF, open reading frame; EIN2, ethylene signaling gene; FLC1, flowering repression gene; CHS, chalcone synthase; Δ12, Δ12-desaturase; AG, agamous; CLV3, clavata 3; AP1, apetala; PAN, periantha; CBL, cystathionine β-lyase; Δ9, Δ9-desaturase; BYDV-Pol, barley yellow dwarf virus RNA-dependant RNA polymerase (ORFs 1 and 2).

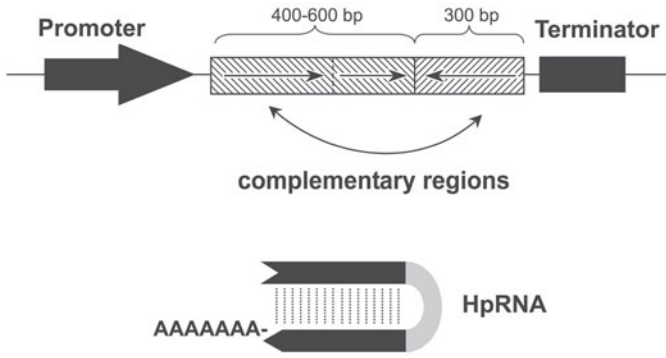


Fig. 1. Conventional hpRNA constructs are made by joining a 400–600-bp target gene sequence to an approx 300 bp fragment from the 3' end of the same sequence in an inverted orientation. Transcripts transcribed from such constructs will have regions of self-complementarity that have the potential to form hpRNA duplexes. An hpRNA construct consists of a sense and an antisense arm separated by a spacer or loop DNA.

vectors. The following parameters are of common consideration for selecting a target gene fragment.

1. Size: gene fragments ranging from 50 bp to 1 kb have been successfully used as targets (*see* **Table 1**). Two factors can influence the choice of length of the fragment. Shorter fragments result in a lower frequency of silencing, and very long hairpins increase the chance of recombination in bacterial host strains. The effectiveness of silencing also appears to be gene-dependent and could reflect accessibility of the target mRNA or the relative abundances of the target mRNA and the hpRNA in cells where the gene is active. We recommend a fragment length of 300–600 bp as a suitable size to maximize the efficiency of silencing obtained.
2. Sequence: both translated as well as untranslated regions (UTRs) have been used with equally good results (*see* **Table 1**). As the mechanism of silencing depends on sequence homology, there is potential for cross-silencing of related mRNA sequences. Where this is not desirable, a region with low sequence similarity to other sequences, such as a 5' or 3' UTR, should be chosen. To reduce cross-silencing, blocks of sequence with identity over 20 bases between the construct and nontarget gene sequences should be avoided.

1.1. Conventional hpRNA Constructs

In their simplest form hpRNA constructs can be made from either the whole or part of the target gene sequence as illustrated in **Fig. 1**. The efficiency of these constructs may not be as high as the hairpins containing an intron ***(5)*** (*see* **Table 1** for details regarding efficacy of such constructs in plants).

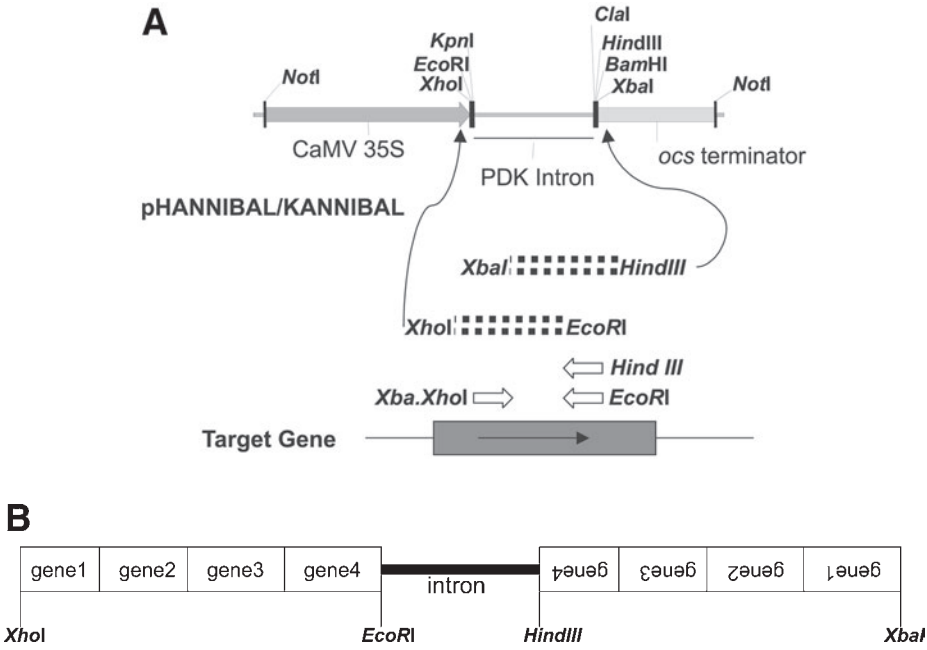


Fig. 2. **(A)** The gene of interest is PCR-amplified with the indicated restriction enzymes appended to the 5' end of the primers and sequentially cloned into similarly cut pHANNIBAL or pKANNIBAL vectors. Cloning into *Xho*I.*Eco*RI.*Kpn*I polylinker gives the sense arm of the hairpin and cloning into *Cla*I.*Hin*dIII.*Bam*HI.*Xba*I polylinker the antisense arm. **(B)** When silencing multiple genes, fragments from various genes are PCR-amplified, stitched together, and the whole cassette is cloned in the sense and antisense orientation into pH/KANNIBAL vectors.

These constructs can be assembled in a primary cloning vector such as pART7 *(8)*, which contains a promoter for constitutive expression in both monocot and dicot plants. Once the assembly of the inverted repeat is complete, it can then be cloned into an appropriate binary vector, such as pART27 *(8)*, for transformation and expression in plants. A 400–600-bp sequence of the target gene is amplified in a PCR. This PCR fragment can then be ligated to an approx 300-bp fragment from the 3' end of the same target gene sequence in an inverted orientation. Transcripts transcribed from such constructs will have regions of self-complementarity that have the potential to form hpRNA duplexes. Thus, the hpRNA construct, in essence, consists of a sense and an antisense arm separated by a spacer or loop DNA. While it is imperative that the inverted repeat part of the construct consists of sequences of the target gene, the spacer region can consist of any DNA fragment (*see* **Note 1**).

1.2. The pHANNNIBAL and pKANNIBAL Vectors

The pHANNIBAL (with ampicillin resistance in bacteria)/pKANNIBAL (with kanamycin resistance in bacteria) system (**Fig. 2A**) is found to work very efficiently and effectively for a number of genes (**Table 1**) and is suitable for silencing a small number of genes, but is laborious when individually silencing a large number of target genes. The construction of each hpRNA construct usually takes around 2 wk. A PCR fragment could be inserted, using conventional restriction enzyme digestion and DNA ligation techniques, in the sense orientation into the *Xho*I.*Eco*RI.*Kpn*I polylinker and in the antisense orientation in the *Cla*I.*Hin*dIII.*Bam*HI.*Xba*I polylinker.

The pKANNIBAL vector is particularly useful, because the PCR fragments from the target gene can be directly cloned, without prior restriction enzyme digestion, into a commercially prepared 3' T-overhang ampicillin-resistant vector, such as pGEM®-T Easy (Promega), and then subcloned into pKANNIBAL using differential antibiotic selection. The *Not*I fragment from pH/KANNIBAL, containing the hpRNA cassette, can then be subcloned into a convenient binary vector such as pART27 (resistance to spectinomycin in bacteria and to kanamycin in plants) and used to transform plants. The pGEM Teasy:pKANNIBAL:pART27 cloning system bypasses the need for purifying DNA fragments from a gel because of the differential antibiotic selection in bacteria.

1.3. The pHELLSGATE Vectors

These vectors were designed as a high-throughput alternative to the pH/KANNIBAL vectors using the commercially available Gateway cloning system (www.invitrogen.com) in which the int/xis system from bacteriophage is modified to allow unidirectional in vitro cloning. As well as allowing directional cloning, the system incorporates a negative selection marker (ccdB) that selects against vectors that have not undergone a recombination reaction, resulting in a high frequency of recovery of recombined plasmids.

The pHELLSGATE vectors contain two recombination cassettes consisting of either attP1-ccdB-attP2 or attR1-ccdB-attR2 in an inverted repeat configuration, such that when gene fragments flanked by the appropriate att sites are recombined with the vector, an ihpRNA-encoding construct is produced **(Fig. 3)**. A series of HELLSGATE vectors are now available **(Fig. 4)**. Constructs in pHELLSGATE4 are generated by a single recombination with an attB-flanked PCR product; however more effective silencing is observed with constructs in pHELLSGATE 8 or 12. In these vectors, the gene fragment is recombined into an intermediate vector, such as pDONR201, before a second recombination into pHELLSGATE8/12. pHELLSGATE12 contains two introns in opposite

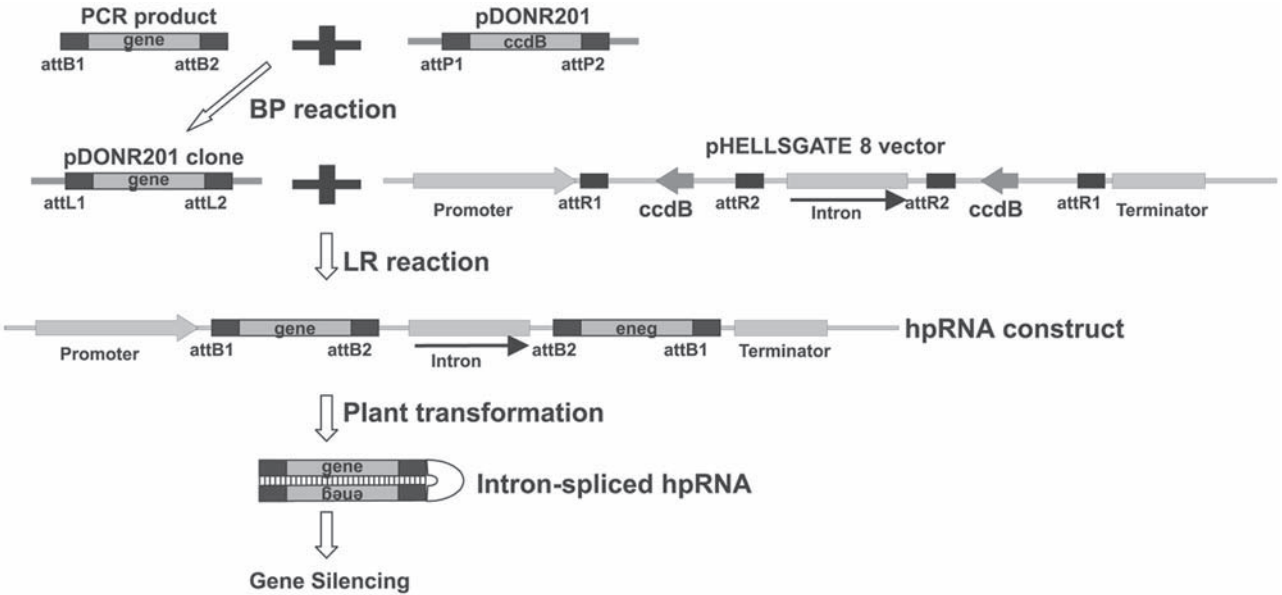


Fig. 3. To clone into pHELLSGATE8, the gene of interest is amplified with primers that have attB1 and attB2 sites appended to the 5' and 3' ends, respectively. The PCR product is directionally recombined into pDNOR201 vector through an in vitro recombination reaction using the enzyme BP clonase. The pDONR201 clones are then recombined into pHELLSGATE8 in a second recombination reaction using an enzyme LR clonase. The resultant plasmid is capable of producing HpRNA in plant cells transformed with it.

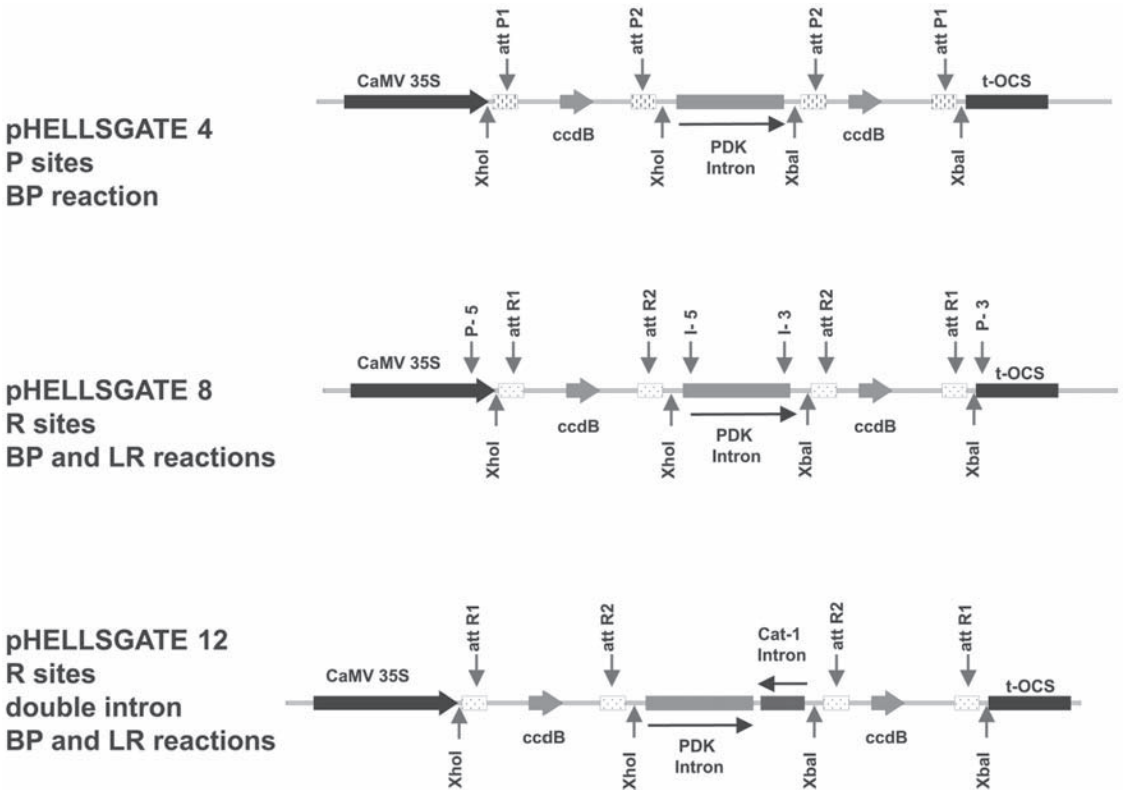


Fig. 4. Constructs in pHELLSGATE4 are generated by a single recombination with an attB-flanked PCR product. In pHELLSGAT8 and 12, the gene fragment is recombined into an intermediate vector, pDONR201, before a second recombination into pHELLSGATE8/12. pHELLSGATE12 contains two introns in opposite orientations, so that the final product of recombination will always contain one spliceable intron.

orientations, so that the final product of recombination will always contain one spliceable intron, thus reducing the number of recombinant plasmids that must be screened to obtain ihpRNA constructs.

2. Materials

2.1. Conventional hpRNA Constructs

1. PCR primers to amplify approx 800 bp of the target sequence with *Eco*RI and *Xho*I appended to the 5' end of the forward and reverse primers, respectively.
2. PCR primers to amplify approx 300–600 bp fragment from the 5' end of the target sequence with *Hin*dIII and *Sma*I appended to the 5' end of the forward and reverse primers, respectively.
3. pART7 ***(8)*** (modify restriction sites on the primers if you are using other vectors).
4. PCR purification kit or columns (Wizard PCR Kit; Promega).
5. 20 ng of DNA template.
6. 10 μ*M* of each primer.
7. 10 m*M* dNTP mixture.
8. 25 m*M* $MgCl_2$.
9. *Taq* DNA polymerase (Applied Biosystems).
10. Buffer containing 50 m*M* KCl, 10 m*M* Tris-HCl, pH 8.3.
11. Appropriate restriction enzymes (from companies such as Promega, MBI Fermentas, etc.).
12. pART27 ***(8)*** or other binary vectors compatible to your plasmid containing 35S promoter.
13. LB medium (liquid and solid) with appropriate antibiotics.

2.2. The pHANNIBAL and pKANNIBAL Vectors

1. Forward primer: 5'-*Xba*I.*Xho*I plus gene specific sequence.
2. Reverse primer: 5'-*Cla*I.*Kpn*I plus gene specific sequence.
3. Vectors such as pGEM Teasy (with resistance to ampicillin in bacteria) to clone the PCR product.
4. pHANNIBAL and pKANNIBAL vectors (www.pi.csiro.au).
5. pART27 vector (www.pi.csiro.au).
6. PCR purification kit or columns (Promega or others).
7. PCR amplification reagents (*Taq* DNA polymerase, buffer, dNTPs, etc.).
8. Appropriate restriction enzymes.
9. LB medium (liquid and solid) with appropriate antibiotics.
10. Primers for sequence verification of hairpin constructs (P-5: 5'-GGGA TGACGCACAATCC-3'; P-3: 5'-GAGCTACACATGCTCAGG-3'; I-5: 5'-ATAATCATACTAATTAACATCAC-3' I-3: 5'-TGATAGATCATGTCA TTGTG-3'.
11. LB plates containing rifampicin (25 mg/L), gentamycin (25 mg/L), and spectinomycin (50 mg/L).

12. Plants to be transformed.
13. 5% Sucrose.
14. Silwet L-77.
15. MS Agar plates containing kanamycin (100 mg/L).

2.3. The pHELLSGATE Vectors

Materials are given for cloning into pHELLSGATE8 vector.

1. Forward primer: attB1-(5'-GGGGACAAGTTTGTACAAAAAAGCAGGCT) plus gene sequence.
2. Reverse primer: attB2-(5'-GGGACCACTTTGTACAAGAAAGCTGGGT) plus gene sequence.
3. AttP1 primer: 5'-GCTAGCATGGATCTCGG.
4. AttP2 primer: 5'-GAGCTGCAGCTGGATGG.
5. BP Clonase, buffer, and proteinase K (Invitrogen; cat. no. 11789013).
6. LR Clonase, buffer, and proteinase K (Invitrogen; cat. no. 11791019).
7. pHELLSGATE 8 (www.pi.csiro.au).
8. pDONR201 (Invitrogen; cat. no. 11798014) (*see* **Note 2**).
9. 30% Polyethylene glycol (PEG), 30 m*M* $MgCl_2$, Tris-EDTA (TE).
10. PCR reagents (as in **Subheading 2.1.**).
11. Water bath at 25°C.
12. LB plates containing kanamycin (50 mg/L).
13. LB plates containing spectinomycin (100 mg/L).

3. Methods

3.1. Conventional hpRNA Constructs

1. Clone the target gene in a sense orientation in the *Xho*I/*Eco*RI sites of the pART7 vector.
2. Set up a standard PCR using 20 ng of DNA template, 0.2 μ*M* of each primer, 200 μM of each nucleotide, 1.5 m*M* $MgCl_2$, and 2.5 U of *Taq* DNA polymerase in 1× buffer. Adjust the reaction vol to 100 μL with water and carry out 30 cycles of amplification using a PCR program consisting of denaturation at 94°C for 1 min, annealing at 55°C for 1 min, and extension at 72°C for 2 min; followed by a further extension at 72°C for 7 min.
3. Clean the PCR product with a purification column.
4. Digest approx 500 ng of the PCR product with *Sma*I and *Hin*dIII.
5. Clean the reaction, resuspend in 10 μL water, and clone into *Sma*I/*Hin*dIII restricted pART7 containing the target gene fragment from **step 1**.
6. Digest a positive clone from **step 5** with *Not*I and clone into a *Not*I-digested pART27 binary vector.

3.2. Cloning into pHANNIBAL and pKANNIBAL Vectors

1. Set up PCR as described in **Subheading 3.1.**, **step 2**.
2. Clean PCR product with a column and digest with *Xho*I and *Kpn*I for sense arm cloning and *Xba*I and *Cla*I for antisense arm cloning.
3. Ligate digested fragments sequentially to *Xho*I/*Kpn*I- and *Xba*I/*Cla*I-digested pHANNIBAL or pKANNIBAL cloning vectors.
4. Clone the *Not*I fragment containing the ihpRNA cassette from pH/KANNIBAL into the *Not*I site of binary vector pART27.
5. For sequence verification (*see* **Note 8**), digest miniprep DNA with *Bgl*II (it cuts once in the pdk intron sequence found in pHANNIBAL, pKANNIBAL, and pHELLSGATE8).
6. Set up two separate PCRs, using P-5 and I-5 primers to amplify the sense arm, and I-3 and P-3 to amplify the antisense arm (the size of the product is 250 bases longer than the insert) (*see* **Fig. 4**).
7. Purify the PCR product and sequence the reactions using the appropriate primers.
8. Transform the hpRNA construct into an *Agrobacterium tumefaciens* strain, such as GV3101, and plate the cells on rifampicin, gentamycin and spectinomycin plates (*see* **Note 9**).
9. Grow liquid cultures of *Agrobacterium* with antibiotics overnight; spin the cultures and resuspend in 2× volume of 5% sucrose and 0.05% Silwet.
10. Transform plants (any *Arabidopsis* ecotype) by the floral dip method; dip them twice, 1 wk apart, collect the seed, and select the transformed plants on kanamycin (100 mg/L).
11. Screen at least 20 independent transformed lines and measure the varying degrees of silencing, either by the severity of the phenotype or by RNA levels (*see* **Note 10**).

3.3. Cloning into pHELLSGATE8 Vector

1. PCR amplify the gene of interest using the forward and reverse primers.
2. Check the PCR products by agarose gel electrophoresis for yield and product size.
3. Purify by diluting the PCR with 3 vol of TE and precipitating with 2 vol of 30% PEG 8000, 30 m*M* $MgCl_2$ (*see* **Note 3**).
4. Collect precipitate by centrifugation at >13,000*g* for 15 min and remove supernatant using a pipet.
5. Resuspend DNA pellet in 1 vol of TE.
6. Set up the BP reaction by mixing: 2 μL BP clonase buffer, 2 μL PCR product, 2 μL (150 ng) pDONR201, and 2 μL TE and 2 μL BP clonase.
7. Incubate at room temperature (25°C) for 1 h.
8. Add 1 μL proteinase K mixture (supplied with BP clonase), incubate for 10 min at 37°C.
9. Use 2 μL to transform *Escherichia coli* DH5α cells (The competent cells should have a transformation efficiency of at least 10^7 colonies/mg plasmid DNA).

10. Plate the transformation mixture on the kanamycin plates.
11. Screen the clones (typically six; *see* **Note 4**) for the insert by restriction digestion with enzymes, such as *Apa*I and *Pst*I, that cut on either side of the insert. Alternatively, PCR amplify the fragment using AttP1 and AttP2 primers.
12. Set up LR reaction by mixing 2 μL LR clonase buffer, 2 μL (100–200 ng) pDONR clone (positive clone from **step 11**), 2 μL (300 ng) pHELLSGATE8 vector, 2 μL TE, and 2 μL LR clonase.
13. Incubate 1–16 h at room temperature (25°C) with longer incubations being better. Treat reaction with 1 μL proteinase K for 10 min at 37°C.
14. Use 2 μL of the reaction mixture to transform DH5α, select colonies on the spectinomycin plates (the plates generally require 24 h incubation at 37°C before colonies are visible).
15. Screen the clones (typically six; *see* **Note 4**) by digesting the miniprep DNA with *Xho*I (sense arm) and *Xba*I (antisense arm) separately (*see* **Note 5**). The size of the fragment should be the size of the insert plus 250 bp (**Fig. 4**).
16. Sequence verify the final clones as in **steps 5–7** of **Subheading 3.2**.
17. Transform plants as in **steps 8–11** of **Subheading 3.2**.

4. Notes

1. If there is an intron adjacent to your target sequence, you could use that as a spacer in designing the hairpin. This may reduce the number of steps.
2. The use of proprietary PCR cleanup columns is not recommended, as the long oligonucleotides cannot be removed from the PCR product.
3. Vectors containing the negative selectable marker ccdB (pHELLSGATE 8,12, and pDONR201) must be maintained in the DB3.1 *E. coli* strain. Competent cells can be purchased from Invitrogen. Alternatively, electrocompetent cells can be prepared using standard methods.
4. The percentage of positive clones obtained in pDONR201 and pHELLSGATE vectors sometimes depends on the gene sequence, which means you may have to screen more than six colonies.
5. The *Xho*I, *Xba*I digestion is not very good on DNA from *Agrobacterium*. Back transformation to DH5α cells, may be necessary.
6. Multiple genes: it has been possible to combine different hpRNA-mediated silenced traits through sexual crossing of relevant transgenic lines ***(11)***. However as the different hpRNA transgenes are inserted at different locations, they will segregate in subsequent generations, thus making the task of stacking modified traits through crossing laborious and time-consuming. This will limit the number of genes that can be combined. An alternative strategy is to use a single hpRNA construct containing inverted repeat of fused multiple target gene sequences (**Fig. 2B**).
7. *See* refs. ***16–18*** for more applications.
8. It is difficult to sequence hpRNA constructs, as the two arms of the hairpin anneal to each other before the primers can anneal to them.
9. Once the hairpin constructs are assembled, they can be stably integrated into plant

genome by plant transformation ***(9)*** or delivered in a transient manner through bombardment or agroinfiltration (*see* **Note 7**) (for review, *see* ref. ***10***). hpRNA silencing is stably inherited up to five generations ***(15).***

10. For designing probes for Northern hybridizations or primers for real-time PCR, use a region in the gene that is not used in the hpRNA construction as, although the target mRNA is degraded, the hpRNA seems to remain intact ***(13)***.

References

1. Hannon, G. J. (2002) RNA interference. *Nature* **418,** 244–251.
2. Gonczy. P., Echeverri, C., Oegema, K., et al. (2000) Functional genomic analysis of cell division in *C-elegans* using RNAi of genes on chromosome III. *Nature* **408,** 331–336.
3. Fraser, A. G., Kamath, R. S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408,** 325–330.
4. Waterhouse, P. M., Graham, M. W., and Wang, M.-B. (1998) Virus resistance and gene silencing in plants is induced by double-stranded RNA. *Proc. Nat. Acad. Sci. USA* **95,** 13959–13964.
5. Smith, N. A., Singh, S. P., Wang, M.-B., Stoutjesdijk, P., Green, A., and Waterhouse, P. M. (2000) Total silencing by intron-spliced hairpin RNAs. *Nature* **407,** 319–320.
6. Wesley, S. V., Helliwell, C. A., Smith, N., et al. (2001) Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J*. **27,** 581–590.
7. Helliwell, C. A., Wesley, S. V., Wielopolska, A. J., and Waterhouse, P. M. (2002) High throughput vectors for efficient gene silencing in plants. *Functional Plant Biol.* **29,** 1217–1225
8. Gleave, A. P. (1992) A versatile binary vector system with a T-DNA organisational structure conducive to efficient integration of cloned DNA into the plant genome. *Plant Mol. Biol*. **20,** 1203–1207.
9. Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana. Plant J*. **16,** 735–743.
10. Horser, C., Abbott, D. C, Wesley, S. V., Smith, N. A., and Waterhouse, P. M. (2001) Gene silencing—principles and application, in *Genetic Engineering*, *Vol. 23*. (Setlow, J., ed.), Kluwer Academic, New York.
11. Liu, Q., Singh, S., and Green, A. (2002) High-stearic and oleic cottonseed oils produced by hairpin RNA-mediated post-transcriptional gene silencing. *Plant Physiol*. **129,** 1732–1743.
12. Chuang, C. F. and Meyerowitz, E. M. (2000) Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proc. Nat. Acad. Sci. USA* **97,** 4985–4990.
13. Levin, J. Z., Framond, A. J., Tuttle, A., Bauer, M. W., and Heifetz, P. B. (2000) Methods of double-stranded RNA-mediated gene inactivation in *Arabidopsis* and

their use to define an essential gene in methionine biosynthesis. *Plant Mol. Biol.* **44,** 759–775.
14. Wang, M.-B., Abbott D. C., and Waterhouse, P. M. (2000) A single copy of a virus-derived transgene encoding hairpin RNA gives immunity to barley yellow dwarf virus. *Mol. Plant Pathol*. **1,** 347–356.
15. Stoutjesdijk, P., Singh, S. P., Liu, Q., Hurlstone, C. J., Waterhouse, P. M., and Green, A. G. (2002) hpRNA-mediated targeting of the *Arabidopsis* FAD2 gene gives highly efficient and stable silencing. *Plant Physiol.* **129,** 1723–1731.
16. Wang, M.-B. and Waterhouse, P. M. (2002) Application of gene silencing in plants. *Curr. Opin. Plant. Biol*. **5,** 146–150.
17. Klink, V. P. and Wolniak, S. M. (2000) The efficacy of RNAi in the study of the plant cytoskeleton. *J. Plant. Growth. Regul*. **19,** 371–384.
18. Schweizer, P., Pokorny, J., Schulze-Lefert, P., and Dudler, R. (2000) Double-stranded RNA interferes with gene function at the single-cell level in cereals. *Plant J*. **24,** 895–903.

17

Virus-Induced Gene Silencing

S. P. Dinesh-Kumar, Radhamani Anandalakshmi, Rajendra Marathe, Michael Schiff, and Yule Liu

Summary

In the postgenomic era, large-scale functional genomic approaches are necessary for converting sequence information into functional information. A para-genetic approach, called virus-induced gene silencing (VIGS), offers a rapid means of gaining insight into gene function in plants. VIGS system could be used to suppress endogenous gene expression by infecting plants with a recombinant virus vector (VIGS vector) carrying host-derived sequence. Here, we describe the use of tobacco rattle virus (TRV)-based VIGS technique to study gene function in *Nicotiana benthamiana* and tomato.

Key Words

gene silencing, VIGS, virus-induced gene silencing, RNAi, TRV-based VIGS vector, functional genomics, tomato, *Nicotiana*

1. Introduction

In recent years, genome sequencing efforts have uncovered large number of open reading frames (ORF). Functions of some of these ORFs could be predicted based on the homology; however, in many cases, ORF sequence alone fails to provide any clue with respect to their function. Therefore, in the postgenome era, functional genomic approaches, like virus-induced gene silencing (VIGS) in plant and RNA interference (RNAi) in animal systems, are promising techniques that will aid in the quick study of functions of unknown genes.

VIGS functions via a posttranscriptional gene silencing (PTGS) mechanism by targeting and degrading RNA in a sequence-specific manner ***(1,2)***. Silenc-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

ing the expression of a gene in a whole plant is one of the ways by which its biological function could be determined. In order to silence a specific gene by VIGS, a recombinant virus vector (VIGS vector) is engineered to carry part of the desired gene sequence. Infection and systemic spreading of this recombinant virus causes specific silencing of the corresponding host gene. Using VIGS, any gene could be silenced in <3 wk.

Several plant viruses have been used to develop VIGS vectors, such as tobacco mosaic virus (TMV) ***(3)***, potato virus X (PVX) ***(4)***, tomato golden mosaic virus (TGMV) ***(5)***, and tobacco rattle virus (TRV) ***(6,7)***. A good VIGS vector should be able to infect the plant and spread rapidly and uniformly, including meristematic regions of the plant. In addition, it should not produce any strong suppressors of silencing. Among the different viruses, TRV exhibits most of these properties. TRV-based vectors have been successfully used to silence genes in *Nicotiana benthamiana* ***(6,7)***, tomato ***(8)***, and potato (Jonathan Jones, Sainsbury laboratory, Norwich, England; personal communication). It infects plants without causing any chlorotic or necrotic symptoms, facilitating easy identification of the VIGS-induced phenotype. TRV infects every cell of the plant and, thereby, induces uniform silencing phenotype. In addition, TRV can also invade and silence genes in the meristems and flowers.

TRV is a bipartite positive sense RNA virus ***(9)***. RNA1 encodes 134- and 194-kDa replicase proteins from the genomic RNA, a 29-kDa movement protein, and 16-kDa cysteine-rich protein from subgenomic RNAs (**Fig. 1**). TRV RNA1 can replicate and move systemically without RNA2. In the Ppk20 strain, RNA2 encodes coat protein from the genomic RNA and two nonstructural proteins from the subgenomic RNAs ***(9)***. To develop TRV-based VIGS vector, cDNA clones of RNA1 and RNA2 were inserted into a T-DNA expression cassette ***(7)***. The cDNAs corresponding to RNA1 and RNA2 were cloned in between duplicated cauliflower mosaic virus (CaMV) 35S promoter and a nopaline synthase (NOS) terminator. In the TRV RNA2 cDNA construct, the nonessential structural genes were replaced with a multiple cloning site (MCS) useful for cloning the target gene sequences for VIGS (**Fig. 1**). To induce VIGS, TRV RNA1 (pTRV1) and RNA2 (pTRV2) containing *Agrobacterium tumefaciens* bacterial cultures are mixed and infiltrated onto the leaves of *N. benthamiana*. Viral RNA synthesized inside the plant cell following *Agrobacterium* infiltration presumably serves as templates for further replication of viral RNA by the RNA-dependent RNA polymerase encoded by RNA1. Systemic infection by the recombinant TRV then brings about VIGS of the targeted plant host sequences.

This paper describes in detail the protocol to perform VIGS assays in *N. benthamiana* and tomato using the TRV vector.

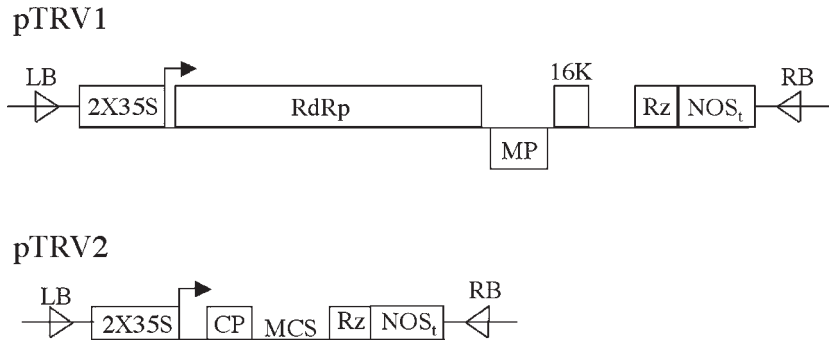


Fig. 1. TRV-based VIGS vectors. TRV cDNA clones were placed in between duplicated CaMV 35S promoter (2X35S) and NOS terminator (NOSt) in a T-DNA vector. RdRp, RNA-dependent RNA polymerase; 16K, 16-kDa cysteine rich protein; MP, movement protein; CP, coat protein; LB and RB, left and right borders of T-DNA, respectively; Rz, self-cleaving ribozyme; MCS, multiple cloning sites.

2. Materials

2.1. VIGS in N. benthamiana

1. *N. benthamiana* seeds.
2. *A. tumefeciens* strain GV2260.
3. GV2260 harboring pTRV1.
4. pTRV2 to clone target gene for VIGS.
5. pTRV2-*NbPDS* as a positive control.
6. 3'–5' Dimethoxy 4'-hydroxy acetophenone (acetosyringone) (200 m*M* stock in dimethyl formamide [DMF]).
7. 2-[*N*-Morpholino] ethane sulfonic acid (MES).
8. 1 *M* $MgCl_2$.
9. Infiltration medium: 10 m*M* $MgCl_2$, 10 m*M* MES, and 200 μ*M* acetosyringone) in sterile water.
10. 1-mL Syringe.

2.2. VIGS in Tomato

1. VF36 or MicroTom seeds.
2. *A. tumefeciens* strain GV3101.
3. GV3101 harboring pTRV1.
4. pTRV2 to clone target gene for VIGS.
5. pTRV2-tomato*PDS* as a positive control.
6. 3'–5' Dimethoxy 4'-hydroxy acetophenone (acetosyringone) (200 m*M* stock in DMF).
7. MES.
8. 1 *M* $MgCl_2$.

9. Infiltration medium: 10 m*M* $MgCl_2$, 10 m*M* MES, and 200 μ*M* acetosyringone) in sterile water.
10. 1 mL Syringe.
11. Artist's airbrush (Model V180; Paasche) connected to a portable air compressor (Campbell Havsfeld).

3. Methods

3.1. VIGS in N. benthamiana

3.1.1. Growing Plants for VIGS

1. Germinate *N. benthamiana* seeds in soil in a pot at 23°–25°C. Cover the pots with Saran® wrap to prevent drying and to provide adequate moisture (*see* **Note 1**).
2. Transplant 2-wk-old seedlings individually into separate pots.
3. Infiltrate plants at the four leaf stage with the *Agrobacterium* cultures (*see* below) (*see* **Note 2**).

3.1.2. Construction of pTRV2 Containing Target Gene for Silencing

1. Select 500–700 bp region of a target gene that needs to be silenced and clone into pTRV2 vector (*see* **Note 3**).
2. Introduce, separately, pTRV1 and pTRV2, carrying the target gene, into *A. tumafeciens* strain GV2260 by electroporation. Select transformants on LB plate containing kanamycin (50 mg/L), rifampicin (25 mg/L), streptomycin (50 mg/L), and carbenicillin (50 mg/L) (*see* **Note 4**).
3. Check the transformants by polymerase chain reaction (PCR) or restriction digestion to confirm the presence of pTRV1 and pTRV2 carrying the target gene.

3.1.3. Infiltration of Agrobacterium into N. benthamiana Leaves

1. Inoculate *Agrobacterium* strain GV2260 containing plasmids pTRV1 and pTRV2 each into 5 mL LB media containing kanamycin (50 mg/L), rifampicin (25 mg/L), streptomycin (50 mg/L), and carbenicillin (50 mg/L). Grow overnight at 28°C.
2. Inoculate the 5-mL overnight cultures individually into fresh 50 mL media containing antibiotics as above. Supplement the media with 10 m*M* MES and 20 μ*M* acetosyringone. Grow the culture overnight at 28°C.
3. Spin down the bacterium at 3000 *g* for 10 min.
4. Resuspend the pellet initially in about 5 mL of infiltration media (10 m*M* $MgCl_2$, 10 m*M* MES, and 200 μ*M* acetosyringone). Dilute with infiltration media to a final OD_{600} of 1.00 (make up approx 20 mL) (*see* **Note 5**).
5. Incubate the culture at room temperature for 3 h (*see* **Note 6**).
6. Mix *Agrobacterium* cultures containing pTRV1 and pTRV2 with target gene in 1:1 ratio.
7. Infiltrate two lower leaves of *N. benthamiana* plants using a 1-mL needleless syringe. A small slit of 0.1 mm could be made on the leaf using a razor blade at the site of infiltration to facilitate introduction of the culture into leaf cells. Place

the opening of the syringe on the slit. Then, using your finger from other side create a pressure to infiltrate the bacterial culture. The penetration of bacterial suspension can be seen clearly as it spreads in the leaf (*see* **Note 7**).
8. Maintain plants at 24°–26°C with adequate light in the growth chamber or conviron.
9. Suppression effect should be seen between 6–10 d (*see* **Note 8**).

3.2. VIGS in Tomato

3.2.1. Growing Plants for VIGS

Germinate tomato seeds in the same way as described above for *N. benthamiana* and transplant into separate pots. Tomato seedlings with two fully developed true leaves are ready for VIGS assays.

3.2.2. Construction of pTRV2 Containing Target Gene for Silencing

1. Clone the target gene sequences into pTRV2 following the strategy described above in **Subheading 3.1**.
2. Introduce, separately, plasmids pTRV1 and pTRV2, carrying the target gene, into *A. tumafeciens* strain GV3101 by electroporation. Select *Agrobacterium* on LB plate containing kanamycin (50 mg/L) and gentamycin (50 mg/L) (*see* **Note 9**).
3. Check transformants by PCR or restriction digestion to confirm the presence of pTRV1 and pTRV2 carrying the target gene in the *Agrobacterium*.

3.2.3. Introduction of Agrobacterium into Tomato Leaves

1. Inoculate *Agrobacterium* strain GV3101, containing pTRV1 and pTRV2 individually, in 5 mL LB media with kanamycin (50 mg/L) and gentamycin (50 mg/L). Incubate the culture at 28°C.
2. Use these 5-mL overnight cultures to inoculate fresh 50 mL LB media containing the same antibiotics, 10 m*M* MES, and 20 μ*M* acetosyringone.
3. Spin down the bacteria on the following day and resuspend the pellet in a solution containing: 10 m*M* $MgCl_2$, 10 m*M* MES, and 200 μ*M* acetosyringone. Adjust the OD_{600} of the culture to 1.5 (*see* **Note 10**).
4. Incubate at room temperature for 3 h.
5. Infiltration: mix pTRV1 and pTRV2 *Agro* in 1:1 ratio; and infiltrate onto two leaves. Make sure that the entire leaf is infiltrated with culture (*see* **Note 11**). Spray: add a pinch of carborundum to the culture. Using an artist airbrush attached to a pressure compressor (set to approx 80 psi), spray the plant from approx 8 in away. Try to spray the underside of each leaf individually, for about 1 s (*see* **Note 12**).
6. After infiltration–spray, allow the plants at 23°–26°C with adequate lighting.
7. Silencing phenotype could be seen around 14 d after infiltration.

4. Notes

1. It is best to germinate seeds in the growth chamber or in the laboratory on a light cart. A greenhouse can be used for this purpose except during hot summer days.
2. Younger plants are better than older plants for VIGS assays.
3. It is possible to use a 300-bp fragment for silencing. With fragments smaller than 300 bp, the silencing effect may be reduced. If the complete sequence of the gene is known; it is better to use its 5' untranslated region for silencing, to avoid suppression of other highly homologous genes.
4. *Agrobacterium* strain GV2260 works best in *N. benthamiana*. Strain GV3101 could also be used. The helper plasmid of GV2260 with *vir* genes carries carbenicillin resistance gene. The chromosomal background is derived from the strain C58C1-RS. This strain is resistant to rifampicin and streptomycin. pTRV2 T-DNA harbor kanamycin resistance gene.
5. It is advisable to make fresh acetosyringone stock. Dissolve acetosyringone in DMF. Lower concentration also works, but higher than 1 OD_{600} may cause necrosis on the infiltrated leaf.
6. Minimum incubation time is 3 h. It is required for the induction of *vir* genes.
7. It is easier to perform infiltration from the underside of the leaf.
8. Always perform pTRV2 alone infiltration as described above as a negative control and pTRV2-*PDS* infiltration as a positive control. Suppression of *PDS* leads to the inhibition of the carotenoid synthesis, causing the plants to exhibit a photobleached phenotype ***(7)***. Photobleaching spreads into upper leaves by day 5. Many upper leaves are completely photobleached by day 10.
9. GV3101 works best for tomato. LBA4404 and GV2260 could be used, but efficiency of silencing is very low.
10. OD_{600} = 1.5 works better for tomato.
11. Success rate using infiltration in tomato is about 30–50%. Therefore, at least 10 individual plants should be infiltrated per construct.
12. Silencing success rate using spray technique is about 90% ***(8)***.

Acknowledgments

We thank the members of S.P. D.-K. laboratory for comments and critical reading of the manuscript. The National Science Foundation Grant DBI-0211872 supports VIGS work in S.P. D.-K.'s laboratory.

References

1. Waterhouse, P. M., Wang, M. B., and Lough, T. (2001) Gene silencing as an adaptive defense against viruses. *Nature* **411,** 834–842.
2. Baulcombe, D. C. (1999) Fast forward genetics based on virus-induced gene silencing. *Curr. Opin. Plant Biol.* **2,** 109–113.
3. Kumagai, M. H., Donson, J., della-Cioppa, G., Harvey, D., Hanley, K., and Grill, L. K. (1995) Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA. *Proc. Natl. Acad. Sci. USA* **92,** 1679–1683.

4. Ruiz, M. T., Voinnet, O., and Baulcombe, D. C. (1998) Initiation and maintenance of virus induced gene silencing. *Plant Cell* **10,** 937–946.
5. Peele, C., Jordan, C. V., Muangsan, N., et al. (2001) Silencing of a meristematic gene using geminivirus-derived vectors. *Plant J.* **27,** 357–366.
6. Ratcliff, F., Martin-Hernandez, A. M., and Baulcombe, D. C. (2001) Technical advance. Tobacco rattle virus as a vector for analysis of gene function by silencing. *Plant J.* **25,** 237–245.
7. Liu, Y., Schiff, M., Marathe, R., and Dinesh-Kumar, S. P. (2002) Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *Plant J.* **30,** 415–429.
8. Liu, Y., Schiff, M., and Dinesh-Kumar, S. P. (2002) Virus induced gene silencing in tomato. *Plant J.* **31,** 777–786.
9. MacFarlane, S. A. (1999) Molecular biology of the tobraviruses. *J. Gen. Virol.* **80,** 2799–2807.

18

Exploring the Potential of Plant RNase P as a Functional Genomics Tool

Dileep K. Pulukkunat, M. L. Stephen Raj, Debasis Pattanayak, Lien B. Lai, and Venkat Gopalan

Summary

As we trek into the uncharted territories of the genomic era, there is an urgency for the development of approaches for assigning functions to the multitude of uncharacterized genes. Although currently available knock-out methodologies could be used for uncovering the function of newly discovered genes, the mixed outcomes in terms of the success of these approaches in down-regulating gene expression necessitate the development of new functional genomics tools. This chapter describes in detail the experimental method for targeted mRNA degradation inside plant cells by enticing the endogenous and ubiquitous RNase P into recognition of specific mRNAs as non-natural substrates.

Key Words

RNase P, EGS, functional genomics, down-regulation of gene expression

1. Introduction

The rapid acquisition of genome sequences lends immediacy to the design and testing of new methods for determining the function of each gene and generating a comprehensive map of locus-phenotype correlates. In plants, reverse genetics approaches have, so far, primarily involved the use of either insertional mutagenesis or antisense and sense suppression ***(1–3)***. The increasing appreciation that RNAs are not mere passive carriers of genetic information, but also involved in various cellular processes in catalytic roles has led to the development of novel customized RNA-based technologies (e.g., ribozymes, RNA interference [RNAi]) for targeted degradation of a cellular mRNA ***(4–7)***. This article describes the underlying concepts and experimental protocols for

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold

exploiting ribonuclease P (RNase P) for targeted disruption of gene expression in plants.

RNase P, which has been identified in all domains of life and is essential for cell viability, catalyzes the hydrolytic reaction that removes the 5'-leader sequences from sixty odd precursor tRNAs (ptRNAs) to form mature tRNAs ***(8–10)***. The RNase P holoenzyme is generally composed of one RNA subunit and one or more protein subunits depending upon the source. Although the RNA subunit is essential for the ptRNA processing activity of all RNase P variants, catalytic activity in the absence of protein in vitro has been established for the bacterial (and select archaeal) versions, which are by definition true ribozymes. There is marked variation in the biochemical composition and function of the various subunits which constitute the RNase P ribonucleoprotein complex in Bacteria, Archaea, and Eukarya ***(8–10)***. Since the RNase P-mediated approach for targeted degradation of mRNAs is dependent only on the substrate recognition properties of RNase P and not on the subunit make-up or the protein-independent catalytic potential of the RNase P RNA subunits, this discussion will focus on the former aspect.

1.1. RNase P-Mediated Inhibition of Gene Expression

The RNase P-based mRNA degradation approach is based on the employment of a single-stranded external guide sequence (EGS) RNA or DNA, either expressed endogenously from a transgene or administered exogenously, which will hybridize to a target mRNA in a sequence-specific manner (**Fig. 1**) ***(11,12)***. The EGS is designed with specific sequence and structural properties, such that upon its binding to the target mRNA, it generates a bimolecular complex which resembles a ptRNA (**Fig. 1A**) and entices endogenous RNase P into cleaving the target mRNA (**Fig. 1B**). The specificity of targeting is derived from Watson-Crick basepairing. Since this EGS mimics nearly three-fourths of a tRNA molecule, it is termed a 3/4 EGS (**Fig. 1B**). As the goal of mRNA degradation is to eliminate synthesis of the encoded protein, the RNase P-mediated cleavage in the mRNA should be positioned either at or immediately downstream of the initiation codon to ensure complete nontranslatability (**Fig. 1B**).

EGS RNAs have been expressed transgenically to inhibit gene expression in mammalian cells. For instance, replication of influenza virus in mouse cells was successfully blocked by expressing 3/4 EGSs targeted against the nucleocapsid and the polymerase genes of this virus ***(13)***. Similarly, expression of an EGS specific for the mRNA of an *N*-methyl-D-aspartate (NMDA) receptor in neuronal cells induced degradation of this mRNA and down-regulated expression of the NMDA receptor as evidenced by the decreased cytotoxicity of select NMDA-receptor ligands ***(14)***. Other studies have established that exogenous RNA- or DNA-based EGSs can be successfully delivered into mammalian cells

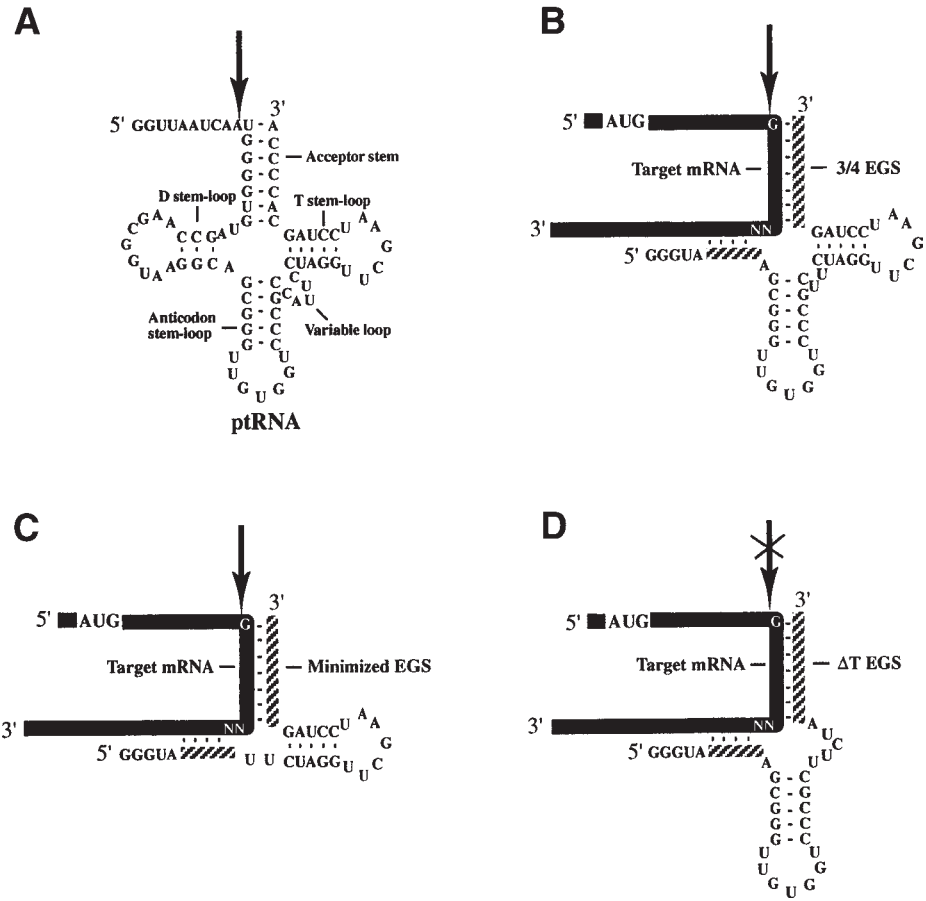


Fig. 1. Rationale for the cleavage of target mRNAs by RNase P *(11,12)*. **(A)** The structure of a ptRNA, a typical substrate for RNase P. The arrow indicates the site of cleavage by RNase P. **(B)** A complex of two RNAs, which are noncovalently bound to form a substrate resembling a ptRNA. RNase P mediates cleavage of a target mRNA (black bar) in the presence of an RNA molecule termed the EGS. In addition to possessing sequences (hatched bars) that are complementary to the target mRNA, the EGS also has the anticodon, variable (minimized), and T stem-loop regions of a typical ptRNA substrate. **(C)** Design of a minimized EGS, in which the anticodon and variable stem-loop regions have been deleted. **(D)** A mutant EGS, which lacks the T stem-loop and serves as a control to measure the contribution of antisense effects to EGS-mediated decrease in gene expression.

and used to elicit degradation of their respective target mRNAs ***(15,16)***. The reader is directed to other reviews for a list of successful applications and variations of this strategy ***(8,12,17)***.

Although evidence of EGS-mediated shut-down of gene expression has been documented in animal cell culture, there are no reports of similar experiments in plant cells. Based on substrate recognition studies in vitro using partially purified rice and maize RNase P, we have recently demonstrated that plant RNase P can also cleave bipartite substrates (with ptRNA-like structures reconstituted from two different RNA molecules) similar to those recognized by human RNase P (**Fig. 1B**) ***(18)***. Encouraged by this finding, we have begun testing the RNase P and EGS-based method in *Arabidopsis* by transgenically expressing EGSs against mRNAs encoding either a reporter protein or a transcription factor involved in stomatal patterning. Since our preliminary results are promising, we furnish details of our approach to promote the use of this method as a potential functional genomics tool.

1.2. Overview of Strategy

In principle, any mRNA of known sequence can be targeted for RNase P-mediated degradation if the following steps are taken. First, the complex secondary structure in the target mRNA can limit the accessibility of EGS binding; therefore, a primary goal is to identify the single-stranded regions in the vicinity of the start codon in the target mRNA. Second, based on the information from RNA structure mapping experiments, a customized EGS is designed with characteristics that render it both specific and efficient in eliciting degradation of the target mRNA. Third, to favor assembly of the EGS-mRNA complex, the EGS needs to be expressed in the appropriate host at a high level, from a strong promoter (either constitutive or regulated). Also, as RNase P is predominantly compartmentalized in the nucleus, which is the site of transcription of nuclear genes, the localization of EGS RNA in this compartment is essential. Lastly, a phenotypic assay is required to evaluate the efficacy of the EGSs in the appropriate transgenic lines. Of course, if an open reading frame (ORF) of unknown function is being targeted, assessment of the target mRNA and protein levels using conventional techniques (such as nuclease protection assay and northern and western blot analysis) would provide evidence for down-regulation of expression of the target gene. In the following sections, we provide information relevant only to the first three aims mentioned above, since methods required to fulfill the last goal are rather generic, and descriptions of the same could be found elsewhere. Guerrier-Takada and Altman, who have pioneered the gene inactivation method that uses bacterial/human RNase P in conjunction with guide sequences, have reviewed in an excellent article the

various steps as well as the technical aspects ***(12)***. Their protocols, in part, form the basis for our experimental design with plant RNase P.

1.3. Selection of the Target Site

Since the EGS binds to the target mRNA on the basis of sequence complementarity (**Fig. 1**), it is important to first identify regions in the target mRNAs that are accessible for binding to the EGS. Several theoretical and experimental approaches are available for identifying regions of low folding potential in the first 200 nucleotides (nts) immediately downstream of the initiator AUG in the target mRNA. Presumably, due to the need for ribosomal entry, the translational initiation site is generally single-stranded and, therefore, expected to be accessible for basepairing to another RNA, such as the EGS. In fact, this premise has thus far been borne out in the various RNase P-based gene targeting studies reported ***(12–14)***, as well as those analyzed in our laboratory. Although RNA secondary structures can be predicted by algorithms such as *mfold* (http://www.bioinfo.rpi.edu/applications/mfold) ***(19)***, these computationally derived structures are based primarily on energy minimization and, therefore, must be interpreted with caution and verified by experimental data. Various enzymatic and chemical probes are available to gain insights into RNA structure ***(20)***. For instance, RNase T1 cleaves preferentially unpaired guanosine residues and aids in mapping single-stranded regions in an RNA. The RNase T1 cleavage pattern of the first 200 nts of the target mRNA can also be validated by the *mfold* prediction.

Studies to date which relied on in vitro RNase T1-based secondary structure mapping data for EGS design have demonstrated that such EGSs function effectively in vivo ***(12)***. Therefore, we believe that this enzymatic probing method, albeit rather simple, might suffice especially since a facile and rapid functional genomics approach should not entail extensive preliminary experimentation. A couple of caveats merit mention. First, a target mRNA might not lend itself to RNase T1-based probing due to a paucity of unpaired Gs. Such a problem can be easily circumvented using several recently developed experimental approaches for mapping sites in an RNA that are accessible for binding to oligonucleotide ***(20–23)***. Second, since RNA structure mapping carried out in vitro on fragments of target mRNA may not reflect its fold in vivo, dimethyl sulfate (DMS), a chemical reporter of adenines and cytosines not involved in Watson-Crick base pairing, could be used to determine target mRNA structure in vivo ***(15,20)***. Chemical probes enjoy an advantage over nucleases in that they are less sensitive to steric hindrance.

1.4. Design of EGS

While designing EGSs for use in plant cells, the following considerations will apply. Once an accessible region is mapped, a guanosine residue 3' to the site of RNase P cleavage in the target mRNA is preferred, since the equivalent position in most ptRNAs, the natural substrates of RNase P, is most often a guanosine. Subsequently, the EGS is designed to be complementary to 11 of 13 nts immediately downstream of this cleavage site in the target mRNA to facilitate formation of 11 bps corresponding to the acceptor and D-stem equivalent of the ptRNA substrate (**Fig. 1B**). Note that a spacer of 2 nts separates the 11 complementary nts into 2 segments of 7 and 4 nts (as in tRNAs) (**Figs. 1A** and **1B**). If there is a choice of single-stranded regions proximal to the AUG in the target mRNA, it is preferable to select the most GC-rich sequence among these regions to ensure strong hydrogen bonding with the EGS. An additional option to enhance target specificity would be to make the EGS complementary to 14 nts, instead of 11, in the target mRNA, such that it generates 7 bps each in the acceptor and D-stem-like regions of the bipartite complex. This design is based on our recent finding that plant RNase P can cleave a modified ptRNA substrate with 7 bps in the D-stem at three-fifths of the rate observed with the wild-type tRNA ***(24)***.

In addition to the regions of complementarity, the EGS also possesses sequences corresponding to the anticodon, variable, and T stem-loop sequences (**Fig. 1B**) ***(8,12)***. Since RNase P is involved in 5' maturation of more than 60 ptRNA substrates in vivo, designing the most effective EGS would entail using as template a ptRNA sequence that is cleaved most efficiently by plant RNase P. Since there are no reports of either a comparative analysis of plant RNase P-mediated cleavage rates of different ptRNAs or an in vitro evolution-based approach to examine which randomized ptRNA sequence might be cleaved with the highest efficiency by plant RNase P, we are currently using in our EGSs the anticodon and T stem-loop sequences present in cyanobacterial $ptRNA^{Gln}$, which is an excellent substrate for *Arabidopsis*, rice, and maize RNase P (**Fig. 1B**) ***(18,24)***.

Recently, our investigations into the substrate recognition properties of plant RNase P have revealed that the anticodon and variable stem-loops are not essential for cleavage ***(24)***. Therefore, we believe that a minimized EGS, in which these domains have been deleted (**Fig. 1C**), might be functional in vivo. In fact, similar experiments have been successful with EGSs and human RNase P in human bladder carcinoma cells ***(16)***. The smaller size of the minimized EGS might be attractive if plant calli or suspension cells are to be bombarded or transfected with chemically synthesized EGSs, which incidentally could be modified to confer nuclease resistance in vivo ***(16)***.

It is conceivable that an EGS-mediated decrease in expression of the target protein stems from antisense effects, because the EGS RNAs are complementary to the target mRNA. Hence, an important control experiment must always be included to address this possibility. Plant RNase P, like its human counterpart, cannot cleave a ptRNA substrate in which the T stem-loop has been deleted ***(24,25)***. This observation is exploited in the design of a mutant EGS (ΔT EGS) (**Fig. 1D**), in which regions of complementarity with the target mRNA are maintained, but the T stem-loop region has been deleted. Since the ΔT EGS can bind the target mRNA, but the resulting complex is not likely to be cleaved by RNase P, any disruption of gene expression observed with the ΔT EGS will indicate the degree of antisense effects. Such control experiments, in studies reported so far, revealed that the inhibition of gene expression observed with the mutant EGS is less than 10% compared to nearly 90% observed with the wild-type EGS ***(13,14)***. Clearly, EGS-mediated disruption of gene expression in vivo is specifically due to RNase P and not attributable to antisense-based effects. These results are perhaps to be expected, since RNase P-based targeted cleavage of an mRNA enjoys the benefit of catalytic turnover, a feature lacking in antisense approaches.

1.5. Expression of EGSs and Assessment of Their Efficacy In Vivo

To test the EGSs in vivo, we have cloned the synthetic gene for various EGSs separately under the control of either the *Arabidopsis* U3 or U6 snRNA pol III promoter (**Fig. 2**). These promoters were chosen because: *(i)* the activity of a pol III promoter is several-fold higher than the activity of a pol II promoter; *(ii)* the regulatory elements in these promoters reside entirely upstream of the coding region of the gene; *(iii)* the promoter of the *Arabidopsis* U6 small nuclear RNA (snRNA) gene has been shown to drive transcription of completely unrelated sequences in chimeric gene constructs; and *(iv)* studies that were successful in using RNase P to disrupt gene expression in mouse and human cells have utilized pol III promoters to express EGSs ***(12–14,26–30)***. The success in these mammalian studies may also be due to the fact that pol III transcripts are localized to the nucleus as is RNase P ***(31)***.

Reverse transcription-polymerase chain reaction (RT-PCR) experiments using as template total RNA isolated from transgenic *Arabidopsis* plants containing the EGS genes indicate that the EGSs are being transcribed in vivo (data not shown). Detailed characterization of these chimeric constructs, in which pol III promoters drive expression of small EGS RNAs, is currently in progress. Meanwhile, we will make available the pCAMBIA1390-derived T-DNA binary vectors, in which either the U3 or U6 promoter has been placed upstream of a multiple cloning site, to facilitate cloning of an EGS of choice (**Fig. 2**).

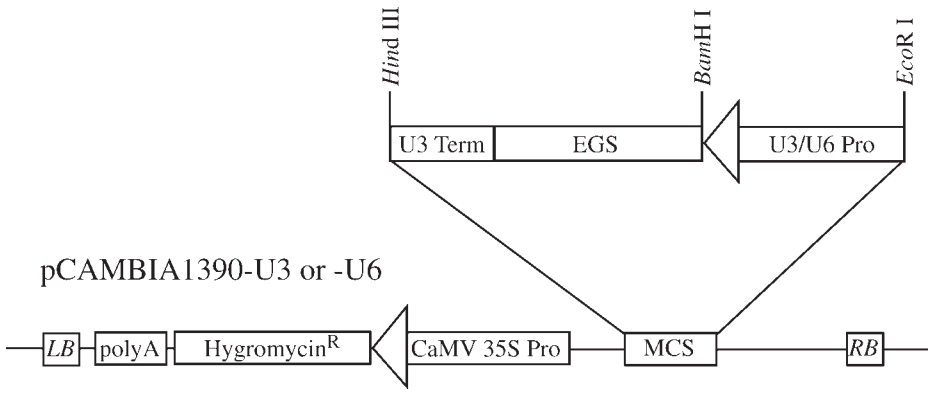


Fig. 2. Map of T-DNA binary vectors containing an *Arabidopsis* U3 or U6 promoter. While two overlapping DNA oligonucleotides were used to construct the promoter fragment corresponding to 77 bps upstream of the transcriptional start site in the U3B gene ***(26,27)***, PCR was used to amplify from *Arabidopsis* genomic DNA the fragment corresponding to 354 bps upstream of the transcriptional initiation site in the U6-29 snRNA gene ***(26)***. These promoter constructs include both the TATA box and an upstream sequence element (USE), bearing the RTCCCACATCG consensus, that are located approx 30 and 60 bps upstream of the transcription initiation site, respectively. These two elements are necessary and sufficient for accurate and efficient transcription of U3 and U6 snRNA genes in dicots (*see* **Note 3**). A cluster of four or more T residues acts as a signal to terminate transcription from these promoters ***(26–29)***.

Once the EGSs are cloned into the binary vector, *Agrobacterium*-mediated transformation of *Arabidopsis* could be used to generate transgenic plants that express the desired EGS against a specific target mRNA ***(32)***. After appropriate screening on an antibiotic-containing medium to ensure the presence of the transgene, the transgenic plants could be evaluated for phenotypic alterations. In addition, total RNA and protein should be isolated from the transgenic plants (expressing both wild-type and ΔT EGSs) for subsequent ribonuclease protection assay (RPA) and Western blot analysis to determine the levels of the target mRNA and protein, respectively. RPAs are preferable over Northern blots due to at least two reasons. First, significantly higher amounts of total RNA can be probed, since the RPA involves hybridization in solution and is not limited by Northern blot constraints, such as the amount of total RNA that could be loaded in a gel. This is an extremely important consideration, if the target mRNA being probed is present at low levels in the host under investigation. Second, the ability to use multiple short probes concomitantly permits a rapid assessment of changes in levels of mRNAs that are closely related (in sequence) to the target mRNA, but that are not expected to be down-regulated by the target

mRNA-specific EGS. Also, EGS-mediated disruption of gene expression could be unambiguously demonstrated if there is an appreciable decrease in the target protein levels in the transgenic plants that express the wild-type EGS (but not the ΔT mutant) against a specific target mRNA.

If our ongoing studies in *Arabidopsis* establish the viability of the RNase P and EGS-based method, the next goal will be to translate it into a reliable high-throughput gene-function discovery approach in plant cells. The fulfillment of this rather daunting objective will be linked at least in part to the development of reliable *in silico* approaches for identifying single-stranded regions in cellular mRNAs and making routine the ability to monitor concomitantly the expression of various cellular mRNAs using microarray analysis. Engineering novel pol III promoters that will lend themselves to temporal and spatial regulation will also facilitate exquisite control over the expression of EGSs and lend additional appeal to the EGS-based strategy as a functional genomics tool.

2. Materials

1. γ-[^{32}P]-ATP (Amersham Pharmacia Biotech).
2. T4 polynucleotide kinase (PNK) (New England Biolabs).
3. RNase T1 (Amersham Pharmacia Biotech).
4. Buffer A: 50 m*M* Tris-HCl, pH 7.5, 100 m*M* NH_4Cl, 10 m*M* $MgCl_2$.
5. Loading dye: 9 *M* urea, 0.05% (w/v) xylene cyanol and 0.05% (w/v) bromophenol blue.
6. Buffer-saturated phenol, pH 7.5 (Life Technologies).
7. Buffer B: 50 m*M* $NaHCO_3$, pH 9.2, 1 m*M* ethylenediaminetetraacetic acid (EDTA) (*see* **Note 1**).
8. Qiagen PCR Purification Kit.

3. Methods

3.1. RNase T1 Mapping to Determine the Accessible Regions in the 5' Portion of a Target mRNA

The region corresponding to the first 200 bps immediately downstream to the initiation codon of the target gene should be subcloned into a suitable vector (such as pBluescript® [Stratagene]) under the control of a T7 RNA polymerase promoter ***(33)***. This clone can then be used for run-off transcription in vitro ***(34)*** and the resulting run-off transcripts labeled at their 5' end using γ-[^{32}P]-ATP and T4 PNK. Subsequent to gel purification, the labeled RNA is subjected to digestion with RNase T1, which cleaves unpaired guanosine residues in a Mg^{2+}-containing buffer. Use of sizing ladders will reveal the exact sites of cleavage by RNase T1 and, thus, identify regions in the RNAs that are single-stranded (*see* **Note 1**).

3.1.1. Digestion of Target mRNA with RNase T1

1. Resuspend approx 25,000 dpm of the 5' end-labeled RNA in 20 µL buffer A and place on ice.
2. Add 1 µL of RNase T1 (50 U/mL) and let the reaction proceed for 30 s.
3. Terminate the reaction by adding 2 µL buffer-saturated phenol. Vortex for 10 s to mix contents.
4. Add sodium acetate and glycogen to a final concentration of 0.3 *M* and 20 µg/mL, respectively.
5. Precipitate RNA with 2 vol of ethanol.
6. Pellet the RNA sample at 18,000*g* for 15 min in a microcentrifuge.
7. Wash the pellet twice with 75% ethanol and dry the sample in a SpeedVac® (Savant Instruments).
8. Resuspend the RNA pellet in 5 µL loading dye.

3.1.2. Alkaline Hydrolysis of Target mRNA

1. Resuspend another portion of the 5' end-labeled RNA pellet (approx 25,000 dpm) in 50 µL of buffer B and aliquot 10 µL each into five 1.5-mL microcentrifuge tubes.
2. Incubate all five tubes at 95°C, each for a different duration: 30, 60, 90, 120, and 150 s. Quench the reactions at the end of the respective incubations by plunging tubes on ice.
3. Pool the contents from the five tubes and add water to bring to a final vol of 100 µL.
4. Precipitate and process the RNA sample as previously described (*see* **Subheading 3.1.1.**).

3.1.3. Gel Electrophoresis

1. Prepare an 8% (w/v) polyacrylamide/7 *M* urea sequencing gel with 1× TBE as the running buffer. The gel could be cast and prerun while performing the steps described in **Subheadings 3.1.1.** and **3.1.2**.
2. Load 2.5 µL (approx 12,000 dpm) each of the partial RNase T1 digest and the alkaline hydrolysis ladder of the target mRNA.
3. Subsequent to high-voltage electrophoresis (50 W for 90 min), dry the gel and obtain an autoradiogram.

3.2. Design and Cloning of the Synthetic Gene Encoding the EGS

Based on the structural mapping data, choose a cleavage site as well as the nucleotides in the target mRNA expected to basepair with the EGS (*see* **Subheading 1.4.**). Note that EGSs could be designed to form either 11 or 14 bps complementary with the target mRNA (*see* **Note 2**). We recommend that the gene for the wild-type and the ΔT EGS (antisense control) be cloned under the control of either *Arabidopsis* U3 or U6 promoters (**Fig. 2**) for experiments in dicots (*see* **Note 3**).

Using a sample target mRNA sequence 5'…NNNGAAAUGA↓*GGGAA GA*UA*GCGC*GCGNNN…3', we illustrate how a synthetic gene encoding either a functional wild-type EGS or a disabled ΔT EGS would be assembled. In the sequence shown above, the initiation codon is in outline, the expected RNase P-mediated cleavage site in the presence of the appropriate EGS is depicted by ↓, and the italicized sequences represent the nucleotides that would potentially basepair with the EGS. The sequence of the wild-type and ΔT EGSs for this target mRNA would be 5'-gggua*GCGC*agcggguugugguccccg cuucuaggguucgaauccuag*UCUUCCC*-3' and 5'gggua*GCGC*agcggguugugguu cccgcuucua*UCUUCCC*-3', respectively. Note that the lowercase letters correspond to cyanobacterial $tRNA^{Gln}$ sequence (with the underlined region denoting the T stem-loop), and the italicized, uppercase letters refer to the nucleotides that are complementary to the target mRNA (correspond to hatched bars in **Fig. 1**).

1. Obtain from a commercial supplier (such as Qiagen or IDT) four DNA oligonucleotides with the following sequences.

   Wild-type EGS forward primer: 5'-CGGGATCCgggta*GCGC*agcgggttgtggtccc gcttctagg-3'
   Wild-type EGS reverse primer: 5'-GGGAAGCTT**AAAAAAAAAGAAAAAAGG AAAGGAC***GGGAAGA*ctaggattcgaacctagaa gcgggacc-3'
   ΔT EGS forward primer: 5'-CGGGATCCgggta*GCGC*agcgggttgtggtcccgctt-3'
   ΔT EGS reverse primer: 5'-GGGAAGCTT**AAAAAAAAAGAAAAAAGGAAA GGAC***GGGAAGA*tagaagcgggaccac-3'

   In the primers shown above, the restriction sites to be employed for subcloning into the binary vectors are double underlined. The key for lowercase and uppercase letters is as described above. The complement of the pol III terminator sequence is shown in bold face in the reverse primers.
2. Anneal the corresponding pair of oligonucleotides and fill in using the Klenow fragment or any suitable DNA polymerase.
3. Use the Qiagen PCR Purification Kit to isolate the double-stranded DNA product generated by the fill-in reaction from unincorporated dNTPs and unused primers.
4. Digest the DNA product with *Hin*dIII and *Bam*HI and ligate it to the binary vector pCAMBIA 1390-U3 (or -U6) already digested with the same restriction enzymes. Transform the *Escherichia coli* cells using ligation mixtures by electroporation (or another suitable method), isolate the plasmid DNA from the transformants, and identify the plasmid bearing the desired construct by restriction digest analysis.
5. If double-stranded DNA sequencing confirms the authenticity of the wild-type and ΔT EGS clones in the binary vectors, transform *Agrobacterium tumefaciens* GV3101 with the appropriate plasmids by electroporation.

6. Proceed to *Agrobacterium*-mediated transformation of *Arabidopsis* using the floral-dip procedure ***(32)***.

4. Notes

1. Partial alkaline hydrolysis of an RNA will generate a ladder corresponding to RNA fragments resulting from strand scission at every phosphodiester linkage in the RNA. If these markers are electrophoresed adjacent to an RNase T1 digest of the same RNA, the spacing between various Gs can be ascertained, and this information, in turn, can be used to establish their exact location in the RNA sequence. To assign sizes to the bands observed in the alkaline hydrolysis ladder, use a couple of oligonucleotides (of known size) that are phosphorylated at their 5' end with γ-[^{32}P]-ATP and T4 PNK. It is preferable to use freshly prepared buffer B for partial alkaline hydrolysis. Alternatively, aliquots that were stored at –20°C immediately after preparation could be used.
2. Although a 3/4 EGS is only complementary to 11 nts in the target mRNA (in the original design), it is important to appreciate that tertiary structure contacts in the mRNA-EGS complex (akin to a ptRNA) will ensure specificity beyond mere complementarity ***(35)***. Nevertheless, as mentioned earlier, EGSs could be tailored to basepair with 14 nts in the target mRNA (7 bps each in the acceptor- and D stem-equivalent regions). A 14-mer sequence would be unique in a transcriptome with a complexity of 2.5×10^8 nts. For the sample target RNA discussed in Subheading 3.2., the sequence (with enhanced specificity) of the wild-type and ΔT EGSs would be 5'-gggua*CGCGCGC*agcggguugugguccccgcu-u<u>cuaggguucga-auccuag</u>*UCUUCCC*-3' and 5'-gggua*CGCGCGC*agcggguugugguccccgcuucua-*UCUUCCC*-3', respectively, with the underlined region corresponding to the T stem-loop of cyanobacterial tRNA^{Gln}.
3. Various studies by Filipowicz and coworkers have provided much of the insights related to the promoters present in plant snRNA genes. Their observation that monocot specific promoter (MSP) elements, in addition to the TATA and USE boxes, are required for pol III transcription of monocot snRNA genes in vivo ***(30)*** implies that pCAMBIA1390-U3 or -U6 might be useful for stable transformation and expression of EGSs only in dicots.

Acknowledgments

Research in V. Gopalan's laboratory is supported by grants from the Ohio Agricultural Research Development Center, Consortium for Plant Biotechnology Research, and Ohio Plant Biotechnology Consortium. L. B. Lai is grateful to Dr. Fred Sack for encouragement and support.

References

1. Martienssen, R. A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl. Acad. Sci. USA* **95,** 2021–2026.
2. Napoli, C., Lemieux, C., and Jorgensen, R. (1990) Introduction of a chimeric chal-

cone synthase gene into petunia results in reversible co-suppression of homologous genes *in trans*. *Plant Cell* **2,** 279–289.
3. van der Kol, A. R., Lenting, P. E., Veenstra, J., et al. (1988) An antisense chalcone synthase gene in transgenic plants inhibits flower pigmentation. *Nature* **333,** 866–869.
4. Tanner, N. K. (1999) Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS Microbiol. Rev.* **23,** 257–275.
5. Storz, G. (2002) An expanding universe of noncoding RNAs. *Science* **296,** 1260–1262.
6. Merlo, A. O., Cowen, N., Delate, T., et al. (1998) Ribozymes targeted to stearoyl-ACP Δ9 desaturase mRNA produce heritable increases of stearic acid in transgenic maize leaves. *Plant Cell* **10,** 1603–1621.
7. Chuang, C.-F. and Meyerowitz, E. M. (2000) Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **97,** 4985–4990.
8. Gopalan, V., Vioque, A., and Altman, S. (2002) RNase P: variations and uses. *J. Biol. Chem.* **277,** 6759–6762.
9. Hall, T. A. and Brown, J. W. (2001) The ribonuclease P family. *Methods Enzymol.* **341,** 56–77.
10. Xiao, S., Scott, F., Fierke, C. A., and Engelke, D. R. (2002) Eukaryotic ribonuclease P: a plurality of ribonucleoprotein enzymes. *Annu. Rev. Biochem.* **71**, 165–189.
11. Forster, A. C. and Altman, S. (1990) External guide sequence for an RNA enzyme. *Science* **249,** 783–786.
12. Guerrier-Takada, C. and Altman, S. (2000) Inactivation of gene expression using ribonuclease P and external guide sequences. *Methods Enzymol.* **313,** 442–456.
13. Plehn-Dujowich, D. and Altman, S. (1998) Effective inhibition of influenza virus production in cultured cells by external guide sequences and ribonuclease P. *Proc. Natl. Acad. Sci. USA* **95,** 7327–7332.
14. Yen, L., Gonzalez-Zulueta, M., Feldman, A., et al. (2001) Reduction of functional N-methyl-D-aspartate receptors in neurons by RNase P-mediated cleavage of the NR1 mRNA. *J. Neurochem.* **76,** 1386–1394.
15. Dunn, W., Trang, P., Khan, U., Zhu, J., and Liu, F. (2001) RNase P-mediated inhibition of cytomegalovirus protease expression and viral DNA encapsidation by oligonucleotide external guide sequences. *Proc. Natl. Acad. Sci. USA* **98,** 14831–14836.
16. Ma, M., Benimetskaya, L., Lebedeva, I., Dignam, J., Takle, G., and Stein, C. A. (2000) Intracellular mRNA cleavage induced through activation of RNase P by nuclease-resistant external guide sequences. *Nat. Biotechnol.* **18,** 58–61.
17. Cobaleda, C. and Sachez-Garcia, I. (2001) RNase P: from biological function to biotechnological applications. *Trends Biotechnol*. **19,** 406–411.
18. Raj, M. L. S., Pulukkunat, D. K., Reckard, J. F., Thomas, G., and Gopalan, V. (2001) Cleavage of bipartite substrates by rice and maize ribonuclease P. Application to degradation of target mRNAs in plants. *Plant Physiol.* **125,** 1187–1190.

19. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148.
20. Moine, H., Ehresmann, B., Ehresmann, C., and Romby, P. (1997) Probing RNA structure and function in solution, in *RNA Structure and Function* (Simons, R. W. and Grunberg-Manago, M., eds.), CSH Laboratory Press, Cold Spring Harbor, NY, pp. 77–116.
21. Allawi, H. T., Dong, F., Ip, H. S., Neri, B. P., and Lyamichev, V. I. (2001) Mapping of RNA accessible sites by extension of random oligonucleotide libraries with reverse transcriptase. *RNA* **7**, 314–327.
22. Pan, W., Devlin, H. F., Kelly, C., Isom, H. C., and Clawson, G. A. (2001) A selection system for identifying accessible sites in target RNAs. *RNA* **7,** 610–620.
23. Amarzguioui, M., Brede, G., Babaie, E., Grotli, M., Sproat, B., and Prydz, H. (2000) Secondary structure prediction and *in vitro* accessibility of mRNA as tools in the selection of target sites for ribozymes. *Nucleic Acids Res.* **28**, 4113–4124.
24. Stiffler, M. A. (2002) Substrate recognition by *Zea mays* RNase P: implications for an RNase P-based functional genomics approach in plants. B.S. (Honors) Thesis, The Ohio State University, Columbus, OH.
25. Yuan, Y. and Altman, S. (1995) Substrate recognition by human RNase P: identification of small, model substrates for the enzyme. *EMBO J.* **14,** 159–168.
26. Waibel, F. and Filipowicz, W. (1990) U6 snRNA genes of Arabidopsis are transcribed by RNA polymerase III but contain the same two upstream promoter elements as RNA polymerase II-transcribed U-snRNA genes. *Nucleic Acids Res.* **18,** 3451–3458.
27. Marshallsay, C., Kiss, T., and Filipowicz, W. (1990) Amplification of plant U3 and U6 snRNA gene sequences using primers specific for an upstream promoter element and conserved intragenic regions. *Nucleic Acids Res.* **18,** 3459–3466.
28. Heard, D. J., Filipowicz, W., Marques, J. P., Palme, K., and Gualberto, J. M. (1995) An upstream U-SnRNA gene-like promoter is required for the transcription of the *Arabidopsis thaliana* 7SL RNA gene. *Nucleic Acids Res.* **23,** 1970.
29. Connelly, S. and Filipowicz, W. (1993) Activity of chimeric U snRNA/mRNA genes in transfected protoplasts of *Nicotiana plumbaginifolia*: U SnRNA 3'-end formation and transcription initiation can occur independently in plants. *Mol. Cell. Biol*. **13,** 6403–6415.
30. Connelly, S., Marshallsay, C., Leader, D., Brown, J. W., and Filipowicz, W. (1994) Small nuclear RNA genes transcribed by either RNA polymerase II or RNA polymerase III in monocot plants share three promoter elements and use a strategy to regulate gene expression different from that used by their dicot plant counterparts. *Mol. Cell. Biol.* **14,** 5910–5919.
31. Bertrand, E., Houser-Scott, F., Kendall, A., Singer, R. H., and Engelke, D. R. (1998) Nucleolar localization of early tRNA processing. *Genes Dev*. **12,** 2463–2468.

32. Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743.
33. Tsai, H.-Y., Lai, L. B., and Gopalan, V. (2002) A modified pBluescript-based vector for facile cloning and transcription of RNAs. *Anal. Biochem.* **303,** 214–217.
34. Vioque, A., Arnez, J., and Altman, S. (1988) Protein-RNA interactions in the RNase P holoenzyme from *Escherichia coli*. *J. Mol. Biol.* **202,** 835–848.
35. Hartmann, R. K., Krupp, G., and Hardt, W.-D. (1995) Towards a new concept of gene inactivation: specific RNA cleavage by endogenous RNase P. *Annu. Rev. Biotechnol.* **1,** 215–265.

19

Maintaining Collections of Mutants for Plant Functional Genomics

Randy Scholl, Martin M. Sachs, and Doreen Ware

Summary

As the plant genomics era progresses and post-genomic functional research rapidly expands, varied genetic resources of unprecedented power and scope are being developed. Partially by the mandate of public funding, these resources are being shared via stock centers and private laboratories. The successful initiation of any new research requires that advantage be taken of these stocks. Information on most plant genomic resources can be obtained through simple yet powerful Web searches, and ordering mechanisms are linked to the information. Hence, locating and obtaining materials is rapid and simple. Currently, available genomic resources are described, and references, links for Web data, and ordering information are also included.

Key Words

T-DNA, transposon, tagging, reverse genetics, flanking sequence

1. Introduction

Mutations by insertion or other traceable genetic events (e.g., deletions) provide powerful tools for cloning genes and genetic analysis. In flowering plants, the targeting of defined areas of the genome is possible through natural properties of transposons, but it has not been possible to target specific genes for modification. However, it is relatively easy to generate near genome saturating numbers of both random T-DNA transformants and transpositions. Consequently, such populations have been developed for cloning and reverse genetics in several species. Publicly available populations will be described, and information on finding and obtaining these will be presented. Emphasis will be placed on *Arabidopsis*, maize, and rice, for which substantial readily available resources exist.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

Several different sources serve as distribution points for genomic resources of plants. Publicly supported stock centers, the *Arabidopsis* Biological Resource Center (ABRC) (supported by the National Science Foundation [NSF]), Ohio State University, and its counterpart organization, the Nottingham *Arabidopsis* Stock Centre (NASC) (supported by Biotechnology and Biological Sciences Research Council [BBSRC]), Nottingham, UK, collect, preserve, and distribute genetic resources of *Arabidopsis*. Similarly, the Maize Genetic Cooperation—Stock Center, United States Department of Agriculture (USDA) (USDA support), Urbana, IL distributes diverse stocks of maize and is the focal point for distribution of many of the maize resources developed in plant genome-funded research. Several complementary facilities maintain and distribute the available resources of rice. It can be expected that public distribution facilities may be developed for various plant species since there is widening research into the genomics of plants. Presently, publicly funded research projects are required to freely share the genetic–molecular resources, which they develop, generally by donation to a stock center (if such exists). For all three of the above species, a few private corporations and individual laboratories also share certain specific resources with the research community. The genomic research materials available from all sources will be described.

2. Resources Available

2.1. Arabidopsis

2.1.1. Types of Insertion Lines–Populations and Their Applications

Random *Agrobacterium* transformants were initially utilized for gene discovery by Feldmann and Marks ***(1,2)***. The procedure was improved by Bechtold and associates ***(3,4)***, so that large-scale forward and reverse genetic exploitation has occurred. Plants produced by transposon and T-DNA insertion constitute the majority of the presently available populations of *Arabidopsis*. Ac/*Ds* and *Spm* of maize are the main transposon systems utilized for insertion analysis on *Arabidopsis*.

A variety of DNA constructions have been employed in the generation of *Arabidopsis* insertional mutants. Reporter constructs in insertion vectors, consisting of coding regions for a visible gene products ***(5)***, are expressed in transformed or transposed vectors when they reside adjacent to an appropriate plant regulatory region. Both β-glucuronidase (GUS) and green fluorescent protein (GFP) have been employed for this approach. Most commonly, promoterless GUS or GFP (promoter trap), "gene trap," or enhancer trap ***(5,6)*** vectors are used. Activation tagging constructs have also been employed ***(7)***.

Arabidopsis T-DNA or transposon populations can be screened for phenotypic effects of any type ***(8)***. Hence, forward genetic screens of large random

T-DNA and transposon populations are often conducted, and the validity of this approach is established (*see* **Note 1**). Numbers of populations are available for this purpose. In addition, these populations are useful for reverse genetics. Sequencing of the plant DNAs, which flank random insertions, has been undertaken by several laboratories, and the corresponding lines have been made available. For other populations, pooled DNA has been isolated so that PCR primers, one from a target a gene and a second from the border of the inserted DNA, can be employed to rapidly assay whether insertions exist in a gene of interest.

The complete genome sequence of *Arabidopsis* represents a powerful adjunct to gene discovery by forward or reverse genetics with T-DNA or transposons. For example, determination of a small segment of chromosomal DNA sequence flanking an insertion should precisely place the insertion site on the genetic map. The complete genomic sequence is available in GenBank®, and copies of the sequence integrated with powerful searching tools reside in *Arabidopsis* databases (the *Arabidopsis* Information Resource [TAIR], http://arabidopsis.org and Munich Information Center for Protein Sequences [MIPS] *Arabidopsis thaliana* Database [MATDB], http://mips.gsf.de/proj/thal/db/index.html).

2.1.2. Available Insertion Resources and How to Locate Them

The resources have been organized below based on type of insertion and intended application. Information on how to utilize the populations, etc., are available in the cited references and/or laboratory Web sites. At some of the Web sites (e.g., notably the site for the Salk insertion lines), robust searching mechanisms are included for locating lines of interest. Most of the relevant data for these resources are integrated into TAIR, so that appropriate searches in this database will find lines of interest. Links for direct ABRC and NASC ordering are provided in TAIR, including availability information for ABRC.

2.1.2.1. T-DNA Populations for Forward Screens

T-DNA populations for the conduct of forward screens are available from ABRC and NASC. The populations listed in the table below can be obtained from the stock centers for minimal fees, which recover part of the costs of reproducing and distributing the material (additional lines from the INRA population may be obtained from NASC):

1. K. Feldmann seed transformed lines ***(1,2)***; 10,000 lines as pools of 10, 20, and 100; T-DNA is a simple insert. Order from ABRC (http://Arabidopsis.org) or NASC (http://Arabidopsis.info).
2. T. Jack lines (http://www.dartmouth.edu/~tjack/); 11,300 lines as pools of 10 and 100; enhancer trap construct. Order from ABRC or NASC.

3. D. Weigel population ***(7,9)***; 23,000 lines as pools of 9–20 and 100; activation tag vector. Order from ABRC or NASC (see Chapter 21 by J. Memelink in this text).
4. Institut National de la Recherche Argonomique (INRA), France population; (http://nasc.nott.ac.uk:8300/Vol2ii/pelletier.html) ***(3,4)***; 20,000 lines available as pools of 20 and 100; enhancer trap construct. Order from ABRC or NASC (see Chapter 14 by M. Rojas-Pierce and P. Springer in this text).
5. C. Koncz lines ***(10)***; 265 individual lines; simple insertion. Order from ABRC or NASC.
6. J. Alonso, W. Crosby, and J. Ecker population ***(8)***; 40,000 lines as pools of 10 and 100; simple insert. Order from ABRC or NASC (see Chapter 11 by J. Alonso and A. Stepanova in this text).
7. W. Scheible and C. Somerville population ***(9)***; 63,000 lines available as pools of 100–350; activation tag vector. Order from ABRC or NASC.
8. *Arabidopsis* Knock-Out Facility (AKF), M. Sussman and R. Amasino population (http//www.biotech.wisc.edu/Arabidopsis) ***(11,12)***; 59,633 lines available as pools of 9 and 225; simple insertion. Order from ABRC or NASC (*see* **Note 2**).
9. R. Bressan and J.-K. Zhu populations (http://www.hort.purdue.edu/hort/people/faculty/bressan.html) ***(13)***; 14,000 lines available as pools of 10 and 100, activation tag vector. Order from ABRC or NASC .
10. B. Bartel population; (http://bioc.rice.edu/~bartel/) ***(14)***; 15,000 lines available as pools of 100; insertions consist of an overexpression library. Order from ABRC or NASC.

Various types of constructs are represented in these materials, including reporter genes and the popular activation tagging types. Selectable markers for identification of plant containing insertions are provided (*see* **Note 3**). In almost all cases, these resources are distributed as pools rather than as single transformant lines. This is due to the extensive labor required for the latter. These pools work well, except when negative screens are required. Screening of single lines also works better when substantial labor in the form of inspection of individual plants is necessary. Multiple insertion sites typically occur for T-DNA transformants, and this must be dealt with as a line is characterized (*see* **Note 4**).

The above populations represent 200,000+ independent transformants and, as such, constitute near-saturation of the genome, except that insertions in any very small target gene(s) are still unlikely to be included.

2.1.2.2. T-DNA Reverse Genetics Resources

The major reverse genetic resources currently available in *Arabidopsis* are organized according to two different assaying principles: *(i)* PCR screens of large populations organized into hierarchical or cross-classified isolated DNA pools; or *(ii)* libraries of individual lines for which flanking DNA sequence of

the T-DNA or transposon insertions have been determined and deposited in a publicly available database(s). The largest source for the former is the collection at the University of Wisconsin, consisting of two populations of 60,000+ lines each. Contact and other information are provided:

1. DNA samples from Feldmann population ***(1,2,15)***; 10,000 lines available as two-dimensionally organized isolated DNA samples based on largest pool size of 1000, small pool of 10, DNA of successively decreasing complexity sent to users. Order from ABRC or NASC.
2. DNA samples from Jack population, 6000; lines available as two-dimensionally organized isolated DNA samples based on largest pool size of 1000, small pool of 10, DNA of successively decreasing complexity sent to users. Order from ABRC or NASC.
3. AKF, U. Wisconsin (http://www.biotech.wisc.edu/Arabidopsis/) ***(11,12)***; screening service accessing 60,000 hierarchical plus 60,000 two-dimensional populations, smallest pool size is 9, PCRs of successive screens conducted on DNA at AKF using users' primers. Consult AKF Web site to use service.
4. Biological Research Center, Institute of Plant Biology, Hungarian Academy of Sciences, Szegad, Hungary; flank sequenced T-DNA lines (http://www.szbk.u-szeged.hu/~arabidop) ***(10)***, collection of flank sequenced T-DNA lines; consult above Web site for a list of tagged genes and to obtain lines.
5. Flanking sequence tags (FST) project, INRA, France (http://flagdb-genoplante-info.infobiogen.fr/projects/fst/DocsIntro/Page_accueil.html) (http://nasc.nott.ac.uk:8300/Vol2ii/pelletier.html) ***(3,4)***; searches for flanking sequences can be completed at the former site; all are seeds are available from NASC with a subset available from ABRC; enhancer trap construct. Order from ABRC or NASC.
6. Sequence indexed Salk lines (http://signal.salk.edu) ***(8)***; 140,000 single lines for which sequence flanking sequence has been conducted, sequence of plant flanking region is published in GenBank; seed lines available from ABRC, and NASC (see Chapter 11).
7. Syngenta/Torrey Mesa Research Institute lines (http://www.tmri.org/pages/collaborations/garlic_files/GarlicDescription.html); 100,000 flank-sequenced single lines private database. The institute was closed in early 2003.

The above resources represent a combined population of 386,000 lines which if all are assayed should provide a high probability of finding an insertion for all but the smallest genes. Further, location of multiple lines having insertions in the same gene is likely (*see* **Note 5**). Antibiotic or herbicide resistance should not be exclusively relied on for identification of insertion plants and especially not for conclusions regarding segregation ratios (*see* **Note 3**).

2.1.2.3. Transposon Resources

All transposon resources of *Arabidopsis* are of two types: *(i)* lines for which transpositions have already been induced and which may be obtained for (mainly) phenotypic screening; and *(ii)* lines with mapped elements that may be induced to move via a cross to a transposase-containing line and generate novel insertions in an adjacent chromosomal region.

2.1.2.3.1. Transposed Lines. For some transposon lines, flanking sequence information is available, and for others, information on reporter expression has been collected. Transposon populations available for reverse genetic screens:

1. Sainsbury Laboratory (SLAT) collection (http://www.jic.bbsrc.ac.uk/sainsbury-lab/jonathan-jones/jjhome.htm) ***(16)***; *Spm* transposed lines having single inserts in each line; effective size of the population is approx 30,000; hierarchically organized pools of DNA may be screened via PCR, smallest pool, 9; also some single sequence tagged lines available with sequences available from the above Web site; order seeds from NASC (*see* **Note 6**).
2. Institute of Molecular Agrobiology transposon tagged lines (http://www.ima.org.sg/) ***(6,17)***. Single sequence tagged *Ds* lines; seeds available from ABRC and NASC.
3. Flank sequenced lines from Cold Spring Harbor Laboratory (CSHL) (http://formaggio.cshl.org/~h-liu/attdb/cgi-perl/blast.cgi) ***(6)***. Large population of flank sequences of *Ds* transposed lines may be searched at the above site and seeds obtained from CSHL or ordered from ABRC and NASC for lines available through the stock centers.

2.1.2.3.2. Transposase Resources Residing at the Stock Centers. The following resources are comprised of lines carrying mapped transposons, so that the tendencies of the transposon to jump to nearby chromosomal locations can be capitalized on to create saturation of knock-outs in the surrounding chromosomal region. Details and references for this approach are included in the maize section, below:

1. Baker transposase lines ***(18)***; mapped *Ds* elements may be mobilized for localized transposition through crosses to *Ac* lines. Order from ABRC or NASC.
2. Fedoroff transposase lines ***(19,20)***; mapped *Ds* elements may be mobilized for localized transposition through crosses to *Ac* lines. Order from ABRC, or NASC.
3. CSHL/Martienssen, Sundaresan lines ***(6)***; genome-wide transposase lines; *Ds* elements may be mobilized for transposition throughout the genome through crosses to *Ac* lines and negative selection against local transpositions. Order from ABRC or NASC.

2.1.3. Additional Reverse Genetics Resource

The targeting of induced local lesions in genomes (TILLING) population (100,000 lines) of Henikoff et al. ***(21,22)***, is accessible via a service facility (http://tilling.fhcrc.org:9366/). They assist in identifying appropriate locations in genes of interest and, subsequently, perform assays with the probes for these locations to identify single-nucleotide substitutions in a heavily ethyl methane sulfonate (EMS)-mutagenized population (see Chapter 13 by S. Henikoff and coworkers in this text). Usually, the screening of a few thousand plants of the population is sufficient to identify more than one EMS mutant of a locus. The seed population for this resource is held at ABRC, so that single lines, in which a mutation has been identified, can be ordered after the screening process is completed at the facility.

2.2. Maize

2.2.1. Types of Insertional Mutants

The existence of transposable genetic elements was first proven by Barbara McClintock ***(23,24)*** when she showed that the *Dissociation* (*Ds*) element moved from one position on chromosome 9S to another and also that this element inserted into the *c1* gene and could again excise from it. This movement was shown by McClintock ***(25)*** to be controlled by an unlinked factor called *Activator* (*Ac*), which itself is a transposable element. Other unstable (or mutable) alleles were known in maize prior to this. The variegated pericap trait was studied by Rollins A. Emerson ***(26)***. This was later shown to be due to the insertion of an element called *modulator of pericarp* (*Mp*) in the pericarp color (*p1*) gene, and the red stripes were due to excision and transposition of this element to another location. The *Mp* element was subsequently shown to be identical to an *Ac* element ***(27)***, and the variegated pericarp trait is encoded by an allele now known as *P1-vv::Ac*. Marcus Rhoades showed that the reference nonpigmented *anthocyaninless* (*a1*) allele was unstable in crosses with black Mexican sweet corn ***(28,29)***, resulting in spotted kernels. In his studies, Rhoades showed that the instability was caused by an element called *Dotted* (*Dt*), which was unlinked to the *a1* gene and, therefore, controlled the *a1-m* instability in a transacting manner. It was not until McClintock discovered the *Ac/Ds* system that it was understood that the *a1-m/Dt* interaction was also due to a transposable element excising from the *a1-m* allele in response to the *Dt* controlling element. This excision of an element, now referred to as a receptor of *Dotted* (*rDt*), allowed for expression of the resulting *A1* revertant allele in progeny cells of the developing aleurone layer, giving the spotted kernel phenotype.

Other transposable elements were subsequently discovered in maize. This includes the *Enhancer/Inhibitor* (*En*) system ***(30)***, which is the same as the *Suppressor/Mutator* (*Spm*) system ***(31)***. The *Mutator* elements ***(32)***, which includes *Cycler* (*Cy*) ***(33)***, are another intensely studied system. Many other transposable element systems are known in maize. These include elements that behave similar to the classic *Ac/Ds* system and include *Bergamo* (*Bg*) ***(34)***, *Factor Cuna* (*Fcu*) ***(35)***, *Mrh* ***(36)***, and *Ubiquitous* (*Uq*) ***(37)***. These all have autonomous (element encodes a transposase, e.g., *Ac*) and receptor/nonautonomous/defective elements (many of which are deletions of autonomous elements, e.g., *Ds*). Other types of transposable elements in maize include *miniature inverted-repeat transposable elements* (MITES) ***(38)***, and retrotransposon-like elements ***(39,40)***.

Tagged gene resources in maize are mostly due to *Ac* and *Mutator* elements. *Ac* has been used because the active element exists in low copy number, and it has a propensity to hop short distances and thus creates new mutations that are tightly linked to its original location ***(41,42)***. The *Mu* system is used because it creates new mutants at a high frequency. *Mutator* elements have a very high propensity for inserting within genes even though the vast majority of the maize genome is intergenic ***(43–45)***.

2.2.2. Resources

2.2.2.1. Maize Genetics Cooperation—Stock Center

The Maize Genetics Cooperation–Stock Center (http://www.uiuc.edu/ph/www/maize) includes classic alleles: from stocks of Barbara McClintock, Peter Peterson, Donald Robertson, etc., and also the Maize Gene Discovery project, Don McCarty's *UniformMu* project, and Hugo Dooner's and Tom Brutnell's *TrAc* projects.

1. Transposed *Ac*s to enable tagging maize genes ***(46)***. *Ac* as a gene-searching engine (http://waksman.rutgers.edu/~dooner/PGRPpage.html). Brutnell: (http://bti.cornell.edu/Brutnell_lab2/Projects/Tagging/BMGG_pro_tagging.html) (regional mutagensis utilizing *Ac* in maize) ***(47)*** (see Chapter 10 by T. Brutnell and L. Conrad in this text).

   These projects are characterizing and mapping transposed *Ac* elements. The goal is that there will be *Ac* elements spread throughout every maize chromosome, such that each could be used as an anchor to tag any closely linked gene. One would choose a Tr*Ac* stock based on the *Ac* element's map distance from a gene of interest. In turn, genes tagged by *Ac* elements can be sequenced.
2. Maize Gene Discovery ***(45):*** engineered *Mutator* element (*RescueMu*) ***(48)***.
   - The *RescueMu* element has pBluescript® as its internal sequence, enabling it and flanking maize DNA to be plasmid-rescued in *Escherichia coli* ***(49)***. This allows for easy isolation and sequencing of genes with *RescueMu* inserts.

Visible mutations caused by *RescueMu* insets, can be correlated with sequence information. Stocks can be ordered from the Maize COOP, based on phenotype and/or sequence (see Chapter 3 by M. Raizadaa in this text).

3. Functional Genomics of Endosperm Development in Maize (http://pgir.rutgers.edu/Functional.html; http://www.endosperm.org). *Mu*-tagged endosperm mutants in a W22 inbred background (*Uniform Mu* population). Other traits are also being screened for. *UniformMu* is a Robertson's *Mutator* population that has been extensively backcrossed into color-converted W22. Each mutant is derived from a pedigreed nonmutant progenitor, and thus, each isolate is assured to be independent.

2.2.2.2. Other Collections and Services

1. Trait Utility System for Corn (TUSC) was developed by Pioneer Hi-Bred International, Inc. ***(50,51)***. TUSC is a reverse genetics tool based on PCR and the *Mutator* transposable element family. If one has sequence information for a gene of interest, mutants of that gene can be found.
2. Maize Targeted Mutagenesis (MTM) project (http://mtm.cshl.org/). A public sector reverse genetics project. The MTM project is a large *Mutator* (*Mu*) population and screening service created by a collaboration between CSHL, Syngenta, and UC Berkeley. The screening service is open to all academic researchers. Insertions into genes of interest are detected by nested PCRs on 3-dimensionally pooled DNA samples. This project is also generating and characterizing visible mutants that are being donated to the Maize Genetics Cooperation—Stock Center.
3. The Cereal Genetics Group, based at the Institute of Arable Crops Research-Long Ashton Research Station, UK, is investigating gene function. In maize, they have produced transposon mutagenesis resources based on Robertson's *Mutator* (*Mu*) transposable elements. These include a classic PCR screen for mutants within known maize genes, a recently developed *Mu* Array screen to rapidly identify mutant plants by hybridization, and a high-throughput screen to identify large numbers of plants with mutations within genes expressed during specific developmental processes. Over 700 amplified *Mu* flanking sequences are now available for searching, and seed for these mutants is available on request (http://www.cerealsdb.uk.net/).

2.3. Rice (Oryza)

Rice, maize, sorghum, wheat, millet, and the other major crop grasses are mankind's most importance source of calories and account for up to 60% of the calories consumed by people in developing world ***(52)***. In 1999, 600 million tons of rice were produced on 155 million hectares globally, 580 million tons of wheat on 215 million hectares, and 600 million tons of maize on 139 million hectares (Food and Agriculture Organization [FAO], [http://www.fao.org]). Ninety-nine percent of rice is consumed directly by humans.

In addition to being a major food source, rice is the rossetta stone of cereal genomics. With the expected completion of the public rice sequence draft in December of 2002, and the recent public release of the two shotgun assemblies from the Syngenta ***(53)*** and the Beijing Genome Institute (BGI) ***(54)***, the potential power of comparative genomics can now be harvested to rapidly accelerate agricultural genomics.

The rice genomic sequence is more than a tool for understanding the biology of a single species. It is a window into the structure and function of genes in the other crop grasses as well ***(54,55)***. Extensive work over the past two decades has shown a remarkably consistent conservation of large segments of linkage groups within rice, maize, sorghum, barley, wheat, rye, sugarcane, and other agriculturally important grasses ***(55–65)***. Based on these and other studies, Dunford et al. ***(66)*** proposed a conceptual framework for collating genetic information on six major grass species by aligning them to 19 rice linkage segments. This work was extended by Gale and Devos ***(67)***, who were able to align the genetic maps of oats, wheat, maize, sorghum, sugar cane, and foxtail millet to just 21 rice linkage groups.

To complement the rice genomic sequence that is becoming available, a number of resources are currently being developed in rice. Many of these resources could and will likely be shared. The insertion resources are mostly not publicly available at the time of writing, although this could change quickly. Hence, the presently available and known potential (in development) community insertional resources will be described below, as will other significant resources for genomic research.

A PCR-based screening service which assays a population of 40,000 T-DNA insertion lines, has been developed by Dr. Gynheung An and associates ***(68,69)*** (information also is at Web site [http://www.postech.ac.kr/life/pfg/]). In this reverse genetics system, users send primers to the facility, where DNA of superpools of plants is subjected to PCR. The reaction products are returned to the user's laboratory for analysis. When an insertion is identified and located to the primary pool, seeds of the respective individual lines are provided to the laboratory for final mutant identification, isolation, and characterization. Use of this resource (e-mail contact for the service: genean@postech.ac.kr) is restricted to academic researchers. Sequencing of the flanking regions of T-DNA insertions of these lines is under way, and the resulting sequence database will be made available for *in silico* identification of lines with insertions of interest and subsequent ordering of the corresponding lines.

Projects to develop technology for and populations of additional insertion lines are in progress. It can be anticipated that such resources will be available from genomics projects funded by U.S. granting agencies. The programs are presently in progress funded by the USDA and NSF and can be viewed at their

Web sites (http://www.reeusda.gov/nri/) and (http://www.nsf.gov/home/grants/grants_awards.htm), respectively.

Very recently, the ability to reliably achieve homologous recombination in rice by T-DNA transformation has been reported ***(69)***. While preliminary reports of similar methods in plants have been previously published, only to be subsequently shown not reliable, the data in this case are especially promising. If this assessment proves correct, rice genomics–genetics will be revolutionized. Further, if the method can be applied to other species, a similar advance in genomics of all plants would occur. The degree to which this approach is valid may already have been at least partially determined when this book reaches publication.

Additional public resources of rice are useful for genomic research. An exceptionally rich base of natural variation exists among the accessions of rice ***(70,71)***. These may be obtained and simply assayed for appropriate phenotypic variation. However, as the genetics of crop stress response, productivity and various quality traits become understood, the appropriate allelic variations for specific genes can be sought with the reasonable expectation that the needed variant may exist. Also, mutant populations from chemical treatments, fast neutrons, and γ rays, respective, have been developed in the IR64 strain at the International Rice Research Institute (IRRI). Genetic variation of the resultant mutant families are being examined, and a database of the phenotypes are recorded in a Web database (http://www.irri.org/genomics/database/IR64.htm). Seeds of the lines will be available under a Material Transfer Agreement. It is planned that 40,000 lines would be included in 2002.

3. Notes

1. While screening T-DNA lines by forward principles has been employed to successfully clone numerous genes and identify loss-of-function alleles, typically no more than one-half of all mutant alleles identified in *Agrobacterium*-transformed populations actually possess a T-DNA. Such alleles cannot be cloned by tagging.
2. The vector utilized in the first Wisconsin AKF population (α set), includes a GUS coding sequence driven by an *APETALA3* promoter. This results in a preponderance of plants expressing GUS in the shoots. In addition, co-supression effects occur, which cause lack–reduction of function effects for floral and shoot characteristics in some lines, especially for flower development.
3. Co-suppression of the plant-selectable marker of a T-DNA initially used to identify transformants may occur in subsequent generations. This effect apparently increases as lines are propagated through additional generations after the initial transformation. This apparently occurs for the kanamycin resistance of the Salk T-DNA collection and may occur for the Basta resistance of the Weigel collection. Since the T-DNA has not been physically altered, selection for its presence can be practiced utilizing appropriate DNA markers in place of resistance to the drug or herbicide.

4. *Agrobacterium*-transformed plants may have T-DNA inserted at a single location, but typically approximately one-half of the transformants have insertions at one or more additional sites. There are approx 1.5 insertions sites per transformed line. Hence, genetic investigations of T-DNA mutants must be conducted in light of this.
5. The large numbers of sequence-tagged insertion lines presently available in *Arabidopsis* allows investigators to collect several lines having independent insertions in a gene of interest. These can subsequently be evaluated, so that the ones that best suit the research needs are used (e.g., ones having simple insertions within the coding sequence, not having additional confounding insertions at additional locations, etc.). Within just the Salk collection, it should often be possible to proceed in this fashion.
6. The SLAT lines each have been engineered so that they carry only a single *Ds* insertion in the absence of *Ac*. This insertion should correspond to the reported flanking sequence in the databases.

References

1. Feldmann, K. and Marks, M. D. (1987) *Agrobacterium*-mediated transformation of germinating seeds of *Arabidopsis thaliana*: a non-tissue culture approach. *Mol. Gen. Genet*. **208,** 1–9.
2. Feldmann, K. A., Marks, M. D., Christianson, M. L., and Quatrano, R. S. (1989) A dwarf mutant of Arabidopsis generated by T-DNA insertion mutagensis. *Science* **243,** 1351–1354.
3. Bechtold, N., Ellis, J., and Pelletier, G. (1993) In planta *Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *C.R. Acad. Sci.* **316,** 1194–1199.
4. Bouchez, D., Camilleri, C., and Caboche M. (1993) A binary vector based on Basta resistance for in planta transformation of *Arabidopsis thaliana*. *C.R. Acad. Sci.* **316,** 1188–1193.
5. Rojas-Pierce, M. and Springer, P. (2003) Gene- and enhancer traps for gene discovery. Ch. 14, this volume.
6. Sundaresan, V., Springer, P., Volpe, T., et al. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9,** 1797–1810.
7. Weigel, D., Ahn, J. H., Blazquez, M. A., et al. (2000) Activation tagging in *Arabidopsis*. *Plant Physiol*. **122,** 1003–1014.
8. Alonso, J. and Stepanova, A. (2003) T-DNA mutagenesis in *Arabidopsis.* Ch. 11, this volume.
9. Memelink, J. (2003) T-DNA activation tagging. Ch. 21, this volume.
10. Koncz, C., Nemeth, G. P., Rédei, G. P., and Schell, J. (1992) T-DNA-mediated insertional mutagenesis. *Plant Mol. Biol*. **20,** 963–969.
11. Krysan, P. J., Young, J. C., Tax, F., and Sussman, M. R. (1996) Identification of transferred DNA insertions within *Arabidopsis* genes involved in signal transduction and ion transport. *Proc. Natl. Acad. Sci. USA* **93,** 8145–8150.

12. Krysan, P. J., Young, J. C., and Sussman, M. R. (1999) T-DNA as an insertional mutagen in *Arabidopsis*. *Plant Cell* **11,** 2283–2290.
13. Ishitani, M., Xiong, L., Stevenson, B., and Zhu, J. K. (1997) Genetic analysis of osmotic and cold stress signal transduction in *Arabidopsis*: interactions and convergence of abscisic acid-dependent and abscisic acid-independent pathways. *Plant Cell* **9,** 1935–1949.
14. LeClere, S. and Bartel, B. (2001) A library of *Arabidopsis* 35S-cDNA lines for identifying novel mutants. *Plant Mol. Biol.* **46,** 695–703.
15. McKinney, E. C., Ali, N., Traut, A., et al. (1995) Sequence based identification of T-DNA insertion mutations in *Arabidopsis*: actin mutants act2-1 and act4-1. *Plant J.* **8,** 613–622.
16. Meissner, R. C., Jin H., Cominelli, E., et al. (1999) Function search in a large transcription factor gene family in *Arabidopsis*: assessing the potential of reverse genetics to identify insertional mutations in R2R3 MYB genes. *Plant Cell* **10,** 1827–1840.
17. Parinov, S., Sevugan, M., Ye, D., Yang, W., Kumaran, M., and Sundaresan, V. (1999) Analysis of flanking sequences from *dissociation* insertion lines: a database for reverse genetics in *Arabidopsis*. *Plant Cell* **11,** 2263–2270.
18. Osborne, B. I., Wirtz, U., and Baker, B. (1995) A system for insertional mutagenesis and chromosomal rearrangement using the *Ds* transposon and Cre-lox. *Plant J.* **7,** 687–701.
19. Fedoroff, N. V. and Smith, D. L. (1993) A versatile system for detecting transposition in *Arabidopsis*. *Plant J.* **3,** 273–289.
20. Smith, D., Yanai, Y., Liu, Y. G., et al. (1996) Characterization and mapping of Ds-GUS-T-DNA lines for targeted insertional mutagenesis. *Plant J.* **10,** 721–732.
21. Colbert, T., Till, B. J., Tompa, R., et al. (2001) High-throughput screening for induced point mutations. *Plant Physiol*. **126,** 480–484.
22. Till, B., Trenton, C., Tompa, R., et al. (2003) High-throughput TILLING for Functional Genomics. Ch. 13, this volume.
23. McClintock, B. (1947) Cytogenetic studies of maize and *Neurospora. Carnegie Inst. Wash. Yearbook* **46,** 146–152.
24. McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA* **36,** 344–355.
25. McClintock, B. (1949) Mutable loci in maize. *Carnegie Inst. Wash. Yearbook* **48,** 142–154.
26. Emerson, R. A. (1914) The inheritance of a recurring somatic variation in variegated ears of maize. *Am. Nat.* **48,** 87–115.
27. Brink, R. A. and Nilan, R. A. (1952) The relation between light variegated and medium variegated pericarp in maize. *Genetics* **37,** 519–544.
28. Rhoades, M. M. (1935) A new aleurone color in maize. *Am. Nat.* **69,** 74–75.
29. Rhoades, M. M. (1938) Effect of the *Dt* gene on the mutability of the a1 allele in maize. *Genetics* **23,** 377–397 .
30. Peterson, P. A. (1953) A mutable pale green locus in maize. *Genetics* **38,** 682–683.

31. McClintock, B. (1951) Mutable loci in maize. *Carnegie Inst. Wash. Yearbook* **50,** 174–181.
32. Robertson, D. S. (1978) Characterization of a *mutator* system in maize. *Mutat. Res.* **51,** 21–28.
33. Schnable, P. S. and Peterson, P. A. (1986) Distribution of genetically active *Cy* transposable elements among diverse maize lines. *Maydica* **31,** 59–82.
34. Salamini, F. (1981) Controlling elements at the opaque-2 locus of maize: their involvement in the origin of spontaneous mutation. *Cold Spring Harbor Symp. Quant. Biol.* **45,** 467–476.
35. Gonella, J. A. and Peterson, P. A. (1977) Controlling elements in a tribal maize from Colombia: *Fcu*, a two-unit system. *Genetics* **85,** 629–645.
36. Rhoades, M. M. and Dempsey, E. (1982) The induction of mutable systems in plants with the high-loss mechanism. *Maize Newsletter* **56,** 21–26.
37. Friedemann, P. and Peterson, P. A. (1982) The *Uq* controlling-element system in maize. *Mol. Gen. Genet.* **187,** 19–29.
38. Wessler, S., Bureau, T. E., and White, S. E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5,** 814–821.
39. SanMiguel, P., Tikhonov, A. P., Jin, Y.-K., et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768.
40. Kumar, A. and Bennetzen, J. L. (1999) Plant retrotransposons. *Ann. Rev. Genet.* **33,** 479–532.
41. Greenblatt, I. M. and Brink, R. A. (1962) Twin mutations in medium variegated pericarp maize. *Genetics* **47,** 489–501.
42. Van Schaik, N. and Brink, R. A. (1959) Transpositions of modulator, a component of the variegated pericarp allele in maize. *Genetics* **44,** 725–738.
43. Bennetzen, J. L. (1996) The mutator transposable element system of maize, in *Transposable Elements* (Saedler, H. and Gierl, A., eds.), Springer-Verlag, New York, pp. 195–229.
44. Chandler, V. L. and Hardeman, K. J. (1992) The *Mu* elements of *Zea mays. Adv. Genet.* **30,** 77–122.
45. Walbot, V. (1991) The *Mutator* transposable element family of maize, in *Current Topics in Genetic Engineering*, *Vol. 13* (Setlow, J. K., ed.), Plenum Press, New York, pp. 1–37.
46. Dooner, H. K., Belachew, A., Burgess, D., Harding, S., Ralston, M., and Ralston, E. (1994) Distribution of unlinked receptor sites for transposed *Ac* elements from the *bz-m2(Ac)* allele in maize. *Genetics* **136,** 261–279.
47. Brutnell, T. and Conrad, L. (2003) Transposon Tagging Using *Activator (Ac)* in Maize. Ch. 10, this volume.
48. Raizada, M. (2003) *RescueMu* Protocol for Maize Functional Genomics. Ch. 3, this volume.
49. Raizada, M., Nan, G.-L., and Walbot. V. (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. *Plant Cell* **13,** 1587–1608.
50. Bensen, R. J., Johal, G .S., Crane, V. C., et al. (1995) Cloning and characterization of the maize *an1* gene. *Plant Cell* **7,** 75–84.

51. Meeley, B. and Briggs, S. P. (1995) Reverse genetics for maize. *Maize Newsletter* **69,** 67–82.
52. FAO Food Balance Sheet (1996) United Nations Food and Agriculture Organization. Rome, Italy.
53. Goff, S. A., Ricke D., Lan, T.-H., et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* **296,** 92–100.
54. Yu, J., Hu, S., Wang, J., et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **296,** 79–92.
55. Whitkus, R., Doebley, J., and Lee, M. (1992) Comparative genome mapping of sorghum and maize. *Genetics* **132,** 119–1130.
56. Melake Berhan A., Hulbert S. H., Butler L. G., and Bennetzen J. L. (1993) Structure and evolution of the genomes of *Sorghum bicolor* and *Zea mays*. *Theor. Appl. Genet.* **86,** 598–604.
57. Grivet, L., D'Hont, A., Dufour, P., Hamon, P., Roques, D., and Glaszmann, J. C. (1994) Comparative mapping of sugar cane with other species within the *Andropogoneae* tribe. *Heredity* **73,** 500–508.
58. Devos, K. M., Beals, J., Nagamura, Y., and Sasaki, T. (1999) *Arabidopsis*-rice: will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Res.* **9,** 825–829.
59. Naranjo, T., Roca, P., Goicoechea, P. G., and Giraldez, R. (1987) Arm homoeology of wheat and rye chromosomes. *Genome* **29,** 873–882.
60. Ahn, S. and Tanksley, S. D. (1993) Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA* **90,** 7980–7984.
61. Ahn, S., Anderson, J. A., Sorrells, M. E., and Tanksley, S. D. (1993) Homoeologous relationships of rice, wheat and maize chromosomes. *Mol. Gen. Genet.* **241,** 483–490.
62. Wilson, W. A., Harrington, S. E., Woodman, W. L., Lee, M., Sorrells, M. E., and McCouch, S. (1999) Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. *Genetics* **153,** 453–473.
63. Van Deynze, A., Nelson, J. C., O'Donoughue, L. S., et al. (1995) Comparative mapping in grasses. Oat relationships. *Mol. Gen. Genet.* **249,** 349–356.
64. Kurata, N., Nagamura, Y., Yamamoto, K., et al. (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. *Nat. Genet.* **8,** 365–372.
65. Devos, K. M., Chao, S., Li, Q. Y., Simonetti, M. C., and Gale, M. (1994) Relationship between chromosome 9 of maize and wheat homoeologous group 7 chromosomes. *Genetics* **138,** 1287–1292.
66. Dunford, R. P., Kurata, N., Laurie, D. A., Money, T. A., Minobe, Y., and Moore G. (1995) Conservation of fine-scale DNA marker order in the genomes of rice and the *Triticeae*. *Nucleic Acids Res.* **23,** 2724–2728.
67. Gale, M. D. and Devos, K. M. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA* **95,** 1971–1974.
68. Jeon, J. and An, G. (2001) Gene tagging in rice: a high throughput system for functional genomics. *Plant Sci.* **161,** 211–219.

69. Terada, R., Urawa, H., Yoshihsige, I., Thugane, K., and Iida, S. (2002) Efficient gene targeting by homologous recombination in rice. *Nat. Biotechnol.* **20,** 1030–1034.
70. Ronald, P. and Leung, H. (2002) The rice genome. The most precious things are not jade and pearls. *Science* **296,** 58–59.
71. Leung, H., Hettel, G. P., and Cantrell, R. P. (2002) International Rice Research Institute: roles and challenges as we enter the genomics era. *Trends Plant Sci*. **7,** 139–142.

20

Vector Construction for Gene Overexpression as a Tool to Elucidate Gene Function

Alan Lloyd

Summary

Gene overexpression as a means to determine plant gene function has been used almost since the first plant transformation protocols became viable. The goal of these experiments, as in classical genetic experiments, is to observe any phenotypic change associated with changing the expression of a gene of interest—in this case overexpression. Any phenotypic changes are interpreted, and the native gene's function is deduced based on the pathways or biochemistries that are altered in the transformants. Overexpression experiments may be particularly suitable in instances when genes are functionally redundant, when a plant species does not have good genetics, or when a knockout mutation is particularly deleterious. This chapter is intended as a general protocol for producing gene overexpression constructs, starting with genomic DNA, RNA, or an isolated clone, for use in plants that are transformable by *Agrobacterium*.

Key Words

overexpression, functional genomics, transformation, CaMV 35S, *Agrobacterium*

1. Introduction

Gene overexpression as a means to determine plant gene function has been used almost since the first plant transformation protocols became viable. For the purposes of this chapter, gene overexpression is defined as using a strong promoter and adjoining upstream activation sequences to drive high level and essentially constitutive transcription of a gene's coding sequence. The desired effect is high steady state mRNA levels and resulting high steady state protein levels. Many of the first experiments using transformed plants were aimed at finding and characterizing strong constitutive promoters, often with the aim of producing highly expressed selectable and scorable markers. As a consequence,

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

much information was available on promoters, such as the cauliflower mosaic virus 19S (CaMV 19S) and 35S (CaMV 35S) transcript promoters and the Agrobacteria opine biosynthetic gene promoters, early in the history of plant transformation. These promoters were and are available to the plant molecular biologist for use in overexpression experiments. The CaMV 35S promoter and its derivatives continue to be the first promoter of choice for off-the-shelf gene overexpression vectors, and many very convenient cassette vectors make construction of gene overexpression plasmids easy ***(1)*** (**Table 1**).

The goal of these experiments, as in classical genetic experiments, is to observe any phenotypic change associated with changing the expression of a gene of interest—in this case, overexpression. Any phenotypic changes are interpreted, and the native gene's function is deduced based on the pathways or biochemistries that are altered in the transformants. This interpretation is preferably done in concert with other molecular data, such as expression patterns and gene similarities. For example, if RNA blot analysis has shown that a gene is undetectable in a tissue or organ, then any phenotypic change due to gene overexpression in that tissue or organ would be viewed with skepticism. However, as always, there are no absolute rules to follow.

There are many reasons to utilize gene overexpression to elucidate gene function. In *Arabidopsis*, maize, petunia, and other genetically tractable plant species, it is generally recognized that the most acceptable way to determine a gene's function is to find a mutation in a gene, by forward or reverse genetics and determine the function by observing a change in phenotype. However, there are many instances where it is not possible to produce the mutant or the mutant gives no differential phenotype. For example, there are many more species that are transformable than have good genetics. Some plants are self-incompatible, so that producing a homozygous recessive individual is difficult, i.e., may take more than one generation. Some plants are polyploid, so that knocking out four or more chromosomal copies of a gene is not practical. Even in diploid plants with good genetics, many genes occur in functionally redundant small or large gene families, so that knocking out one gene produces no observable change. Lastly, some genes are essential, so that gene knock-outs are not recovered as viable plants. In all of these cases, classic genetic analysis may not be practical or possible, while other molecular tools, such as gene overexpression, are practical.

There are also some disadvantages to using gene overexpression to deduce gene function, and those experiments are often criticized for various shortcomings. The major criticism is related to the fact that the gene will almost certainly be expressed out of its normal context. For example, if a gene is expressed at the wrong time or place in development, the wrong phase of the cell cycle, or constitutively instead of being induced by some environmental or

Table 1
Selected Overexpression Cassette Vectors for use with *Agrobacterium*

Vector[a]	Bacterial Selection (mg/L) Same selection for *E. coli* and Agro unless stated otherwise.	Plant Selection (mg/L) Concentration range given. This varies for each species.	Notes	Reference
pBI121	Kanamycin (50)	Kanamycin (25–300)	Used for gene overexpression by excising the GUS gene and replacing it with the gene of interest.	***(38,39)***
pBE2113-GUS	Kanamycin (50)	Kanamycin (25–300)	Derivative of and used like pBI121. Contains multiple CaMV35S enhancers and a translational enhancer for higher level expression.	***(40)***
pGREEN	Kanamycin (50) (pSOUP-Tetracycline [15])	Kanamycin (25–300) Hygromycin (10–50) Bialophos (2–10) Sulfadiazine (20–100)	The pGREEN vectors are small high copy vectors that replicate in Agro with the helper plasmid pSOUP. There are several selectable and scoreable markers and different promoters available. Highly versatile, but may be more complicated to use than other cassettes. See ref. and Web site (http://www.pGreen.ac.uk).	***(41)***
pBINAR	Kanamycin (50)	Kanamycin (25–300)	Derivative of pBIN19 ***(42)*** with simple CaMV35S overexpression cassette.	***(43)***
pKYLX71	Tetracycline (15) in *E. coli*; Kanamycin (50) or Tetracycline (15) in Agro.	Kanamycin (25–300)	Derivative of pGA472 ***(44)*** with simple CaMV35S overexpression cassette.	***(32,45)***

[a]Many other vectors are available.

spatial cue, an altered phenotype may be interpreted inappropriately. In this case, it often helps to have some knowledge of the gene expression pattern to rule out a gene's role in tissues or times where it is not expressed. In addition, many gene overexpression experiments are performed with some knowledge of what is expected to change. Often, several genes in a gene family may be overexpressed where the function of a subset of the members is known. Genes may be overexpressed that have been identified as up- or down-regulated in some specified way during development or after treatment of some sort. In these cases, altered phenotypes can be interpreted with an educated eye and with a modicum of restraint. In the end, overexpression phenotypes appear to be just as reliable indicators of gene function as phenotypes due to other forms of genetic manipulation. In fact, classic rules about assigning epistatic relationships based on mutation genetics have recently been called into question, due mainly to the effects of redundancy ***(2)***. This problem is confounded by the finding that two-thirds of the genes in the *Arabidopsis* genome, and presumably other plant species, are represented by at least one homolog, due to duplication events ***(3)***. Careful interpretation of overexpression phenotypes can often clear up a messy redundancy problem (*see* **Note 1**).

Another criticism of overexpression experiments may be that the transgene is causing a phenotype change, due to insertion in an unrelated gene, and the insertion–knock-out is actually the cause of the change. Producing multiple independent transformants with unlinked insertion positions circumvents this criticism. If most or all of the transformants give a similar phenotype, one can assume it is due to the overexpressed gene. Producing multiple transformants is also desirable to obtain a set of lines with varying levels of transgene expression. Often, a set of 10 transformants will have as much as 100-fold expression level differences and produce mild to severe phenotype changes based on expression levels. This expression level difference, with the same construct, is most often ascribed to insertion position or chromatin context effects. Variations in expression levels can also be due to multiple tandem or unlinked copies of the transgene producing different gene copy numbers.

A protocol is given below that outlines a simple way to produce a vector designed to overexpress a plant gene when the DNA is integrated into the plant genome. The protocol includes methods for producing a genomic or cDNA gene overexpression construct. It is not possible to give a transformation protocol appropriate for of all or even most plant species or a vector appropriate for all plants. The vectors I have listed here are binary vectors for use with *Agrobacterium tumefaciens*-based transformation protocols. *Agrobacterium*-based transformation is presently used with a majority of transformed plants, although some very important species, such as maize and wheat, commonly use free DNA delivery systems, such as particle bombardment or electroporation

with purified plasmid DNA. Transformation protocols vary widely in complexity and degree of technical competence needed. The simplest protocols are for whole plant transformation, such as for *Arabidopsis*, in which no tissue culture experience is required. Other species are less simple, like tobacco and petunia, which require a couple of different tissue culture media. Others are fairly difficult and require substantially more time, space, and resource commitments, such as cotton. However, any species that can be transformed by *Agrobacterium* should work with the vectors listed here. Some other species require transformation methods, such as particle bombardment with DNA-coated bullets. The species that require free DNA delivery transformation are generally more technically demanding. The Agro-binary vectors should work with these species also, without the Agro, but generally, alternate high copy vectors are used. However, the CaMV 35S promoter appears to work in virtually all vascular plant species for gene overexpression. A partial list of species that can be transformed with *Agrobacterium* and in which the CaMV 35S promoter has been used successfully follows: *Arabidopsis* ***(4,5)***, tomato ***(6)***, tobacco ***(7)***, eggplant ***(8)***, quaking aspen ***(9)***, *Allium cepa*, onions and shallots ***(10)***, *Manihot esculenta*, cassava ***(11)***, *Catharanthus* cells ***(12)***, *Cucumis melo* and various squashes and melons ***(13)***, *Lavendula*, lavender ***(14)***, *Malus*, apple ***(15)***, *Asparagus* ***(16)***, *Daucus carota*, carrot ***(17)***, *Lupinus* ***(18)***, *Dyospiros*, persimmon ***(19)***, *Pisum*, pea ***(20)***, *Eucalyptus* ***(21)***, and *Pinus*, a conifer ***(22)***.

Other species where the CaMV 35S promoter works, but Agro is not used for transformation, include: *Zea mays*, corn ***(23)***, although recent literature indicates *Agrobacterium* may be used in the future for maize ***(24)***, it is also interesting to note that the inclusion of an intron leads to increased steady-state levels of mRNAs in maize ***(25)***, so that introns are routinely included in maize overexpression constructs, *Phaseolus*, bean ***(26)***, Arachis hypogea, peanut ***(27)***, Papaya ***(28)***, *Picea*, spruce, a conifer ***(29)***, *Marchantia*, liverwort, a nonvascular plant ***(30)***.

This chapter is intended as a protocol for producing gene overexpression constructs. As mentioned earlier, there are many vectors available for use in gene overexpression experiments, and **Table 1** outlines several vectors containing plant overexpression cassettes that have been fairly heavily used with a variety of species.

2. Materials

2.1. Template Preparation: Genomic DNA Extraction (see Note 2)

1. 2× Hexadecyltrimethylammonium bromide (CTAB) buffer: 2% CTAB (Sigma), 1.4 *M* NaCl, 100 m*M* Tris-HCl, pH 8.0 (from 1 *M* stock), 20 m*M* ethylenediamine tetraacetic acid (EDTA) (from 0.5 *M* stock), 1% polyvinylpyrrolidone, 0.2% β-mercaptoethanol.

2. Chloroform:isoamyl alcohol (CIA) (25:1).
3. TE, pH 8.0: 10 m*M* Tris-HCl, 1 m*M* EDTA, pH 8.0.
4. Isopropanol, ice-cold.
5. 70% Ethanol, ice-cold.
6. 60°C Water bath or heating block.
7. Eppendorf® tubes and Eppendorf pellet pestle (VWR Scientific).
8. 50–100 mg Fresh plant tissue.

2.2. Template Preparation: RNA Isolation Using Trizol™ Reagent

1. Trizol reagent (Life Technologies). Caution, toxic, contains phenol and guanidine isothiocyanate. Best used in fume hood.
2. RNase-free water. Add 0.01% (v/v) diethylpyrocarbonate (DEPC) (Sigma) to distilled water (dH_2O) in baked glass bottles. Keep overnight in fume hood and autoclave.
3. Isopropanol.
4. 75% Ethanol made with RNase-free water in RNase-free Falcon® tube.
5. Chloroform.
6. 50–100 mg Fresh plant tissue expressing the gene of interest.

2.3. Template Preparation: First Strand cDNA Synthesis Using SuperScript™ Reverse Transcriptase

It is most convenient to purchase this as a kit from Life Technologies.

1. SuperScript II reverse transcription (RT) (50 U/μL) (Life Technologies).
2. 10× RT buffer: 200 m*M* Tris-HCl, pH 8.4, 500 m*M* KCl.
3. 25 m*M* $MgCl_2$.
4. 0.1 *M* Dithiothreitol (DTT).
5. 10 m*M* dNTP mixture (10 m*M* each dATP, dCTP, dGTP, dTTP).
6. $Oligo(dT)_{12–18}$ (0.5 μg/μL).
7. RNASEOUT™ Recombinant Ribonuclease Inhibitor (40 U/μL) (Life Technologies).
8. RNaseH.
9. RNase-free water.
10. 37°C, 42°C, 65°C, and 70°C Heating blocks or water baths.
11. Up to 5 μg RNA from **Subheading 3.2**.

2.4. Amplifying and Cloning Gene into Vector

1. Template DNA 1 μL genomic DNA (from **Subheading 3.1.**) or 2 μL first strand cDNA (from **Subheading 3.3.**) or a plasmid containing the gene of interest (*see* **Note 2**).
2. Polymerase chain reaction (PCR) primers with restriction sites, diluted to 500 nM (*see* **Note 3**).
3. *Pfu* polymerase (2.5 U/μL) (Stratagene).

4. 10× *Pfu* polymerase buffer (supplied by manufacturer, if buffer does not contain Mg^{++}, add one-tenth vol of 20 m*M* $MgSO_4$).
5. 2 m*M* dNTP mixture (2 m*M* each dATP, dCTP, dGTP, dTTP).
6. Phenol:chloroform:isoamyl alcohol (PCI) 25:24:1, (v:v:v) pH 8.0.
7. Chloroform.
8. 3 *M* Sodium acetate, pH 5.2.
9. 95% Ethanol.
10. Ice-cold 70% ethanol.
11. Vector pKYLX71 (**Table 1**) (*see* **Note 4**).
12. Restriction enzymes and 10× buffer supplied by vendor.
13. Agarose gel.
14. T4 DNA ligase (Life Technologies).
15. 10× Ligase buffer (supplied by manufacturer).
16. Competent *Escherichia coli* (*see* **Note 5**).
17. Competent *Agrobacterium* (*see* **Note 5**).
18. Selection plates, LB plates with appropriate antibiotic (*see* **Table 1**).

3. Methods

3.1. Genomic DNA Extraction

1. Grind 50–100 mg fresh tissue with pellet pestle in an Eppendorf tube. Grind by hand or with a pestle mounted in an electric drill.
2. Add 50 µL 60°C 2× CTAB buffer and grind again.
3. Add 400 µL 60°C 2× CTAB buffer, mix, and place at 60°C for 30 min to 1 h.
4. Add 800 µL CIA (preferably in a fume hood) and invert several times.
5. Spin for 5 min at 14,000 rpm in a microfuge.
6. Remove aqueous phase (upper) layer to new tube, avoiding interphase.
7. Add 800 µL CIA to new tube and invert several times.
8. Spin for 5 min at 14,000 rpm in a microfuge.
9. Remove aqueous phase to new tube.
10. Add 400 µL ice-cold isopropanol, invert several times, and keep at room temperature for 15–30 min (optional to keep at –20°C for longer period).
11. Spin for 5 min at 14,000 rpm in a microfuge and carefully remove supernatant.
12. Wash pellet with cold 70% ethanol and dry pellet.
13. Resuspend pellet in 50–100 µL TE. Use 1 µL in PCRs.

3.2. Template Preparation: RNA Isolation

1. Place 50–100 mg of fresh plant tissue in an autoclaved Eppendorf tube.
2. Add 1 mL Trizol reagent and homogenize with an Eppendorf pellet pestle. If the tissue is hard or fibrous, it can be powdered in N_2(l) prior to adding Trizol reagent and homogenizing.
3. Incubate the homogenate at room temperature for 5 min and centrifuge at 10,000 rpm for 15 min in a microfuge. Transfer the clear supernatant to a new tube.
4. Add 200 µL chloroform and shake vigorously for 2 min. Centrifuge as above.

5. Remove the upper aqueous phase to a new tube. Add 500 μL isopropyl alcohol and keep at room temperature for 10 min. Centrifuge as above. RNA pellet may be gel-like.
6. Add 1 mL of 75% ethanol to wash. Shake or vortex mix and centrifuge at 5000 rpm for 5 min. Carefully remove the ethanol by pouring or pipeting.
7. Air-dry the pellet for 10 min. Do not over-dry the pellet.
8. Resuspend in 100 μL RNase-free water. Determine the concentration by optical density ($OD_{260} \times 40$ μg/mL × dilution factor).

3.3. Template Preparation: First Strand cDNA Synthesis Using SuperScript™ RT

1. Add up to 5 μg of RNA from above isolation in up to 8 μL vol.
2. Add 1 μL oligo(dT), 1 μL dNTP mixture, and RNase-free water to a total of 10 μL. Mix and place at 65°C for 5 min and on ice for 1 min.
3. Add 2 μL 10× RT buffer, 4 μL 25 m*M* $MgCl_2$, 2 μL 0.1 *M* DTT, 1 μL ribonuclease inhibitor. Mix and place at 42°C for 2 min.
4. Add 1 μL SuperScript II RT and keep at 42°C for 50 min. Place at 70°C for 15 min to terminate the reaction. Place on ice until use in the amplification step or freeze at –20° or –80°C for more permanent storage. We routinely go back to this stock as a source of cDNA for many cloning experiments. It is optional to treat with 1 μL of RNaseH at 37°C for 20 min prior to chilling or freezing. We routinely use 2 μL of this first strand cDNA as template for gene amplification.

3.4. Amplification and Cloning Gene into Vector

1. Design and order primers (*see* **Note 3**).
2. Resuspend primers in double-distilled water (ddH_2O) and dilute to 500 nM concentration (10×).
3. Use an Eppendorf tube appropriate to the thermal cycler. For a 50-μL PCR, add: 5 μL each primer, 5 μL 10× PCR buffer; if buffer contains no Mg^{++}, 5 μL 20 m*M* $MgSO_4$; 5 μL dNTPs, template DNA (1 μL genomic DNA or 2 μL first strand cDNA), ddH_2O to 49 μL, 1 μL (2.5 U) of *Pfu* polymerase (or any proofreading thermostable DNA polymerase, add last to avoid primer degradation). If a thermal cycler with a heated lid is not used, add a drop of mineral oil to cover the reaction.
4. Perform PCR with program appropriate to the length of expected product. For example, the following program would generally be suitable for a 2000-bp product: 5 min at 95°C; (30 s at 55°C, 2 min at 72°C, 30 s at 95°C) repeat 25 times; then 5 min at 72°C. Increase the 72°C incubation by 1 min for every 1000 bp.
5. Check 3–5 μL of product on an agarose gel. Estimate the product concentration by comparing the fluorescent intensity of the product to the intensity of fragments of known mass in a molecular weight marker such as BstEII-restricted λ (New England Biolabs).
6. Add 50 μL PCI and vortex mix (*see* **Note 6**). Centrifuge at top speed in a microfuge for 5 min and move the top aqueous layer to a new tube. Avoid the

interface. Add 50 µL chloroform and centrifuge and recover top aqueous phase as above.

7. Add one-tenth vol 3 *M* sodium acetate and 2.5 vol 95% ethanol. Incubate at –70°C for 30 min or –20°C overnight. Pellet the product at top speed in a microfuge for 15 min, wash with 200 µL ice-cold 75% ethanol, dry pellet, and resuspend in 10 µL TE (*see* **Note 7**).
8. Set up restriction digest with buffer supplied by the enzyme vendor. Digest 200 ng of fragment and 200 ng of vector with the same restriction enzymes in separate tubes at the proper temperature for 2 h (2 µL 10× buffer, 200 ng DNA, water to 19 µL, 1 µL enzyme). If two enzymes that generate incompatible overhangs are used at each end of the gene and vector fragments, the vector will not be able to reclose. Therefore, it is unnecessary to treat the vector with phosphatase. Phenol–chloroform extract and ethanol-precipitate and resuspend in 10 µL TE as in **steps 5** and **6**.
9. Set up two ligation reactions, one with the insert and one without. Add 3 µL vector, 3 µL insert, 1 µL 10× T4 DNA ligase buffer, 3 µL water, 1 µL T4 DNA ligase (approx 400 U). Make a second ligation without the insert and make up the vol difference with water. Ligate at room temperature for 2 h.
10. Transform 1 µL of each ligation into a highly competent *E. coli* strain (*see* **Note 5**). Allow to recover in LB for 1 h at 37°C, pellet the *E. coli*, pour off the supernatant, and resuspend the pellet by vortex mixing in the remaining drip of LB. Plate the entire pellet on LB plates with the appropriate antibiotic selection (**Table 1**) and grow overnight at 37°C. If the cloning worked properly, there should be at least five-fold more colonies on the plate with insert than on the one without.
11. Pick individual colonies with a single sterile toothpick and patch the colony onto a new plate and start a 3-mL liquid culture at the same time. Be sure that the tube and plate patch are numbered consistently. It is common to grow 10–12 colonies from a cloning for analysis. Grow the cultures overnight at 37°C.
12. Isolate plasmid DNA by standard method (*see* **Note 8**). Verify the proper clone by digesting one-tenth of each miniprep with the same enzymes used to prepare the vector and gene insert and separating the fragments by agarose gel electrophoresis. The only two products should be the size of the original vector and the cloned insert.
13. After a vector construct is verified, transform *Agrobacterium* GV3101 pMP90 ***(31)*** by electroporation (*see* **Note 5**). Allow to recover in 1 mL LB at 28°–30°C. Pellet and plate on the appropriate selection (**Table 1**). Grow at 28°–30°C for 2 d. These colonies are the end product that will be used in a plant transformation protocol.
14. In order to check the clones in Agro, we do the same DNA miniprep we would do for *E. coli*. The construct DNA isolated from Agro can be checked by restriction digest as for *E. coli*, or the miniprep DNA can be retransformed back into *E. coli*. The plasmid structure is then verified by restriction digest analysis of DNA isolated from *E. coli*.

3.5. Analysis of Plant Transformants

1. A good goal is to produce at least 10 independent plant transformants. The majority of these transformants should show the same qualitative phenotypic changes. In this way, one can make some generalizations about any phenotypic changes that are observed.
2. Analysis of the transformants is done at many levels, and what is done is often dependent on what plant species is being transformed. In species that have a short life cycle and that readily make many seeds, the segregation and stability of the T-DNA can be assayed by selecting for the dominant selectable marker in the progeny. A 3:1 ratio of resistant to sensitive (tested by χ^2 analysis) indicates a single locus, but higher ratios for 2 or 3 loci are common. Aberrant ratios substantially less than 3:1 are also occasionally observed. We do not pursue these. Southern blots with careful copy number reconstructions can be used to determine T-DNA copy number if necessary. It is common to find multiple tandem copies at a single locus.
3. Analysis of the expression of the transgene is most often done at the mRNA level. RNA (Northern) blots can be used to demonstrate overexpression. Because the 5' and 3' untranslated regions (UTRs) were not included in the transgene constructs described here, the transgene can often be detected as a shorter (and more abundant) mRNA than the endogenous copy. If necessary, the endogenous copy can be detected by using a hybridization probe that is specific for one of the UTR regions not in the transgene. Inclusion of a wild-type untransformed control will allow identification of the transgene and the endogenous gene. In rare cases, a gene intended for overexpression will lead to RNA-mediated pre- or posttranscriptional gene silencing. This complicated subject is beyond the scope of this chapter. However, this type of silencing would be detectable by an mRNA reduction in the silenced plants, if an antibody to the protein is available that can be used in Western blots to assess the protein expression levels in untransformed and transformed plants.
4. Analysis of the phenotypic changes caused by the overexpressed gene is more complicated. Any change could be specific to cells, tissues, time in the life cycle, responses to stimuli, etc. Small, fecund species with short life cycles are usually characterized in later generations, often after plants homozygous for the T-DNA are isolated, while large slow woody species may have to be characterized as the primary transformant, and other species may be intermediate. What to expect.
   a. Too much of something: this is the simplest and perhaps most desirable outcome. Changes might be more enzymatic activity, more pigment, an increased response to stimulus, more of a cell type, earlier or later something when compared to control plants. This will depend on whether the gene encodes a biosynthetic enzyme, a receptor, a positive or negative transcriptional regulator, a cytoskeletal element, etc. In addition, if a new or synergistic phenotype is observed when two genes are co-overexpressed, it is likely that these genes and/or proteins interact in some way.
   b. Gene does nothing: this could be for one of several reasons. The mRNA or the

protein are not stable and cannot be overexpressed in a simple way. The pathway is fully saturated for the activity of this gene product and producing more gives no enhancement. The transforming construct has some mutation or problem that renders it inactive.

c. Gene results in dominant negative that looks the same as a gene knock-out. Two possible reasons are RNA interference and pathway disruption by the protein. RNA interference, commonly called co-suppression, is complicated and mentioned above. However, sometimes a truly overexpressed protein will knock out a pathway even though it is not a negative regulator. This may be the result of "squelching" phenomena, which sometimes occur with proteins that work in complexes ***(32)***. For example, when excess DNA binding proteins bind to a regulatory sequence or to complex partners in an inappropriate context, this can result in the squelching of downstream target genes–pathways.

4. Notes

1. My attempts to find a case of overexpression data leading a plant biologist down the wrong pathway have failed. An example of heterologous ectopic overexpression experiments predicting gene function is that of the maize R gene overexpression ***(33)***, correctly predicting the existence and function of the endogenous *Arabidopsis* bHLH genes, GL3 ***(34)***, Enhancer of GL3 (EGL) (Lloyd and coworkers, unpublished), and TT8 ***(35)***. Endogenous myb genes also seem to maintain specificity when overexpressed, although when expressed in nontarget tissues, they can take the place of similar mybs like the root-expressed, WER, replacing GL1 in the shoot ***(36)***. These and other experiments underscore the potential value of overexpression experiments to produce dominant gene activities for phenotypic analysis and for assigning gene function.
2. When we want to overexpress a gene for which there is no available cDNA template, we most often start with the genomic locus. We amplify the genomic locus from a genomic DNA preparation of our species of choice with little more upstream or downstream sequence past the start and stop codons, respectively. We then clone the genomic PCR product into either a plant transformation vector such as pKYLX71 or pBluescript® (Stratagene) after restriction digest of sites included in the primers. Alternatively, we use a TA cloning vector (Invitrogen) (*see* **Note 7**) to clone the whole product. The resulting clones are sequenced to verify that the correct product was cloned. If pBluescript or TA cloning was used, the product is then subcloned into the plant expression cassette similar to how the PCR product would be directly cloned.

   If a full-length cDNA is available in a plasmid, we will use that as the template for the same primers we would have used above, with the same considerations. If no cDNA exists, it is sometimes easy to produce one that includes the sequence from the start to the stop. A protocol for that is included here. The advantage to using a cDNA clone over genomic is that introns are excluded. The clone is shorter and sometimes (not always), we find we get a more severe phenotype with the

cDNA. Although we have not done a systematic study, we assume that this is due to more net steady-state protein being produced. In any case, we have not yet observed a case where the genomic clone was expressed better than the cDNA.

3. Primers can be ordered from many sources including Invitrogen, Life Technologies, or Integrated DNA Technologies. There are software programs to aid in primer design. However, there often is not much design freedom when creating primers that span the start or stop codons of a gene and that also include convenient restriction sites for subsequent cloning. Most often, I place a convenient restriction site just upstream of the start codon. The sequence around the start codon can be modified to match the Kozak consensus, but in practice, I do not bother to modify the native start context. For example in the following primer, the restriction sites, *Hin*dIII and *Xba*I, are bold, and the start and stop codons are italic. The bases upstream of the sites are to stabilize the ends of the PCR product to aid in restriction digestion, and the Xs are the gene-specific sequences immediately downstream of the start and upstream of the stop: start primer, GGGG**AAGCTT***ATG*XXXXXXXXXXXXXXXXXXX; stop primer, GGGG**TCTAGA***TTA*XXXXXXXXXXXXXXXXX. Keep in mind that the primers are written 5' to 3', and the stop codon primer is the opposite strand. If TA cloning is to be used, the bases upstream of the start and stop codons can be eliminated. I have used primers such as these to amplify and clone genes directly into pKYLX71 on numerous occasions with good result.

   In practice, the restriction sites added to the end of the PCR product should not be present in the gene to be cloned. This is where it becomes an advantage to have several expression cassettes available with different sets of restriction sites.

   Generally, the gene-specific part of the primers, which includes the start or stop codon and the Xs, should have noncomplementary 3' ends and a calculated melting temperature (T_m) of about 55°C. For oligonucleotides shorter than 25 bases, counting just the gene-specific part in the above oligonucleotides, the approximate T_m is commonly calculated with the following formula: T_m (°C) = 2(Number of As + Ts) + 4(Number of Gs + Cs).

4. The *Agrobacterium* vectors in **Table 1** have all been heavily used to transform plants and overexpress genes. For plants that are more commonly transformed by free DNA delivery, smaller high copy vectors are used. The overexpression plasmid is sometimes co-transformed with a second plasmid that contains the selectable marker, rather than placing both on a single plasmid. The steps to construct such an overexpression plasmid would be the same as for the Agro vectors, except that the vector is not transferred to *Agrobacterium*.

5. There are many protocols for transformation of *E. coli* and *Agrobacterium* by free DNA. We use essentially the same protocol for both species. Electroporation competent cells are made by growing 100–500 mL of cells in LB without selection to OD_{600} 0.8–1. Chill cells on ice and keep cold. Pellet the cells at 5000*g* for 5 min. Resuspend in one-half vol ice-cold ddH_2O by shaking vigorously. Pellet and resuspend as before. Pellet as above and resuspend in one-tenth vol ice-cold sterile 10% glycerol. Pellet as above (the pellet may not be as tight, so be careful

when pouring off the supernatant) and resuspend in one-one hundredth of the original vol ice-cold sterile 10% glycerol. Make 90-µL aliquots of cells in small Eppendorf tubes, freeze in (l)N_2, and store indefinitely at –80°C.

Place 40 µL of the competent cells and 1 µL of the ligation for *E. coli* or 1.0 µL of the DNA miniprep for *Agrobacterium* in a cold Eppendorf tube. Place all of this mixture in a cold 0.2-cm electroporation cuvette (Bio-Rad) and electroporate at 2.5 kV on an *E. coli* Pulser (Bio-Rad). Add 1 mL LB to the cuvette, replace cuvette lid, and hold it on tightly while shaking to mix the cells and LB. Pour the mixture into a culture tube and allow to recover at 30°C (Agro) or 37°C (*E. coli*) for 1 h. Pellet and plate on selection as in the protocol. *E. coli* colonies should be visible overnight, and *Agrobacterium* colonies should be visible in 2 d.

6. Clean-up of PCR product. Phenol extraction and ethanol precipitation is usually adequate prior to restriction digestion or ligation. However, many easy-to-use products are available for DNA fragment clean-up. After PCR or restriction digestion, the product or vector can be cleaned-up with a product like Microcon®-PCR Centrifugal Filter Devices (Millipore). In this case, the product is recovered in 20 µL of TE. These are easy and safe devices to use. Follow the manufacturer's recommendations.
7. We often clone PCR products directly into high copy vectors using the TOPO® TA Cloning system (Invitrogen), which allows one to bypass both cleanup and restriction digestion of the PCR product. It is advisable to sequence the cloned PCR product to verify that there are no mutations and the TA vectors facilitate this. The only disadvantage to cloning into other vectors before the plant transformation vectors is that the product will have to be subcloned later into the Agro binary vectors. We have also cloned and sequenced directly from the *Agrobacterium* vectors. If the PCR product is to be cloned into a vector with a 3' T overhang, the product must have a 3' A overhang. *Taq* DNA polymerase will leave such an overhang. If *Pfu* or another proofreading polymerase is used to amplify the product, *Taq* must be added later to add the A overhang. Follow the manufacturer's suggested protocols.
8. There are many options for making miniprep DNA from *E. coli*. These include very easy kits from commercial suppliers like Qiagen or Promega or standard alkaline lysis preps ***(37)***. Most binary vectors are low copy, so we usually extract DNA from 3 mL of culture and resuspend the DNA in a final vol of 50 µL. Five microliters of this is sufficient to see restriction digest bands separated on an agarose gel. One microliter of this is more than enough to transform *Agrobacterium*.

Acknowledgments

Work in the Lloyd laboratory has been supported by the Hermann Frasch Foundation, The Texas Higher Education Coordinating Board, and the National Science Foundation.

References

1. Benfey, P. N., Ren, L., and Chua, N. H. (1990) Combinatorial and synergistic properties of CaMV 35S enhancer subdomains. *EMBO J.* **9,** 1685–1696.
2. Martienssen, R. and Irish, V. (1999) Copying out our ABCs, the role of gene redundancy in interpreting genetic hierarchies. *Trends Genet.* **15,** 435–437.
3. Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* **15,** 2114–2117.
4. Bechtold, N., Ellis, J., and Pelletier, G. (1993) In planta *Agrobacterium*-mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *C.R. Acad. Sci. Paris* **316,** 1194–1199.
5. Lloyd, A. M., Barnason, A. R., Rogers, S. G., Byrne, M. C., Fraley, R. T., and Horsch, R. B. (1986) Transformation of *Arabidopsis thaliana* with *Agrobacterium tumefaciens*. *Science* **234,** 464–466.
6. Janssen, B.-J., Lund, L., and Sinha, N. (1998) Overexpression of a homeobox gene, LeT6, reveals indeterminate features in the tomato compound leaf. *Plant Physiol.* **117,** 771–786.
7. Qin, X. and Zeevaart, J. A. D. (2002) Overexpression of a 9-cis epoxycarotenoid dioxygenase gene in *Nicotiana plumbaginifolia* increases abscisic acid and phaseic acid levels and enhances drought tolerance. *Plant Physiol.* **128,** 544–551.
8. Donzella, G., Spena, A., and Rotino, G. L. (2000) Transgenic parthenocarpic eggplants: superior germplasm for increased winter production *Mol. Breed.* **6,** 79–86.
9. Tsai, C.-J., Popko, J. L., Mielke, M. R., Hu, W.-J., Podila, G. K., and Chiang, V. L. (1998) Suppression of O-methyltransferase gene by homologous sense transgene in quaking aspen causes red-brown wood phenotypes. *Plant Physiol.* **117,** 101–112.
10. Zheng, S.-J., Khrustaleva, L., Henken, B., et al. (2001) *Agrobacterium tumefaciens*-mediated transformation of *Allium cepa* L.: the production of transgenic onions and shallots. *Mol. Breed.* **7,** 101–115.
11. Zhang, P., Potrykus, I., and Puonti-Kaerlas, J. (2000) Efficient production of transgenic cassava using negative and positive selection. *Transgenic Res.* **9,** 405–415.
12. Whitmer, S., Canel, C., Hallard, D., Goncalves, C., and Verpoorte, R. (1998) Influence of precursor availability on alkaloid accumulation by transgenic cell line of *Catharanthus roseus*. *Plant Physiol.* **116,** 853–857.
13. Bordas, M., Montesinos, C., Dabauza, M., et al. (1997) Transfer of the yeast salt tolerance gene HAL1 to *Cucumis melo* L. cultivars and in vitro evaluation of salt tolerance. *Transgenic Res.* **6,** 41–50.
14. Dronne, S., Moja, S., Jullien, F., Berger, F., and Caissard, J.-C. (1999) *Agrobacterium*-mediated transformation of lavandin (Lavandula x intermedia Emeric ex Loiseleur). *Transgenic Res.* **8,** 335–347.
15. Bolar, J. P., Norelli, J. L., Harman, G. E., Brown, S. K., and Aldwinckle H. S. (2001) Synergistic activity of endochitinase and exochitinase from *Trichoderma*

atroviride (*T. harzianum*) against the pathogenic fungus (*Venturia inaequalis*) in transgenic apple plants. *Transgenic Res.* **10,** 533–543.
16. Limanton-Grevet, A. and Jullien, M. (2001). *Agrobacterium*-mediated transformation of *Asparagus officinalis* L.: molecular and genetic analysis of transgenic plants. *Mol. Breed.* **7,** 141–150.
17. Hardegger, M. and Sturm, A. (1998) Transformation and regeneration of carrot (*Daucus carota* L.). *Mol. Breed.* **4,** 119–127.
18. Pigeaire, A., Abernethy, D., Smith, P. M., et al. (1997) Transformation of a grain legume (*Lupinus angustifolius* L.) via *Agrobacterium tumefaciens*-mediated gene transfer to shoot apices. *Mol. Breed.* **3,** 341–349.
19. Gao, M., Sakamoto, A., Miura, K., Murata, N., Sugiura, A., and Tao, R. (2000) Transformation of Japanese persimmon (*Diospyros kaki* Thunb.) with a bacterial gene for choline oxidase. *Mol. Breed.* **6,** 501–510.
20. Anna Nadolska-Orczyk, A. and Orczyk, W. (2000) Study of the factors influencing *Agrobacterium*-mediated transformation of pea (*Pisum sativum* L.) *Mol. Breed.* **6,** 185–194.
21. Harcourt, R. L., Kyozuka, J., Floyd, R. B., et al. (2000) Insect- and herbicide-resistant transgenic eucalypts. *Mol. Breed.* **6,** 307–315.
22. Levee, V., Garin, E., Klimaszewska, K., and Seguin, A. (1999) Stable genetic transformation of white pine (*Pinus strobus* L.) after cocultivation of embryogenic tissues with *Agrobacterium tumefaciens*. *Mol. Breed.* **5,** 429–440.
23. Petolino, J. F., Young, S., Hopkins, N., et al. (2000) Expression of murine adenosine deaminase (ADA) in transgenic maize. *Transgenic Res.* **9,** 1–9.
24. Bronwyn, R. F., Shou, H., Chikwamba, R. K., et al. (2002) *Agrobacterium tumefaciens*-mediated transformation of maize embryos using a standard binary vector system. *Plant Physiol.* **129,** 13–22.
25. Luehrsen, K. R. and Walbot, V. (1991) Intron enhancement of gene expression and the splicing efficiency of introns in maize cells. *Mol. Gen. Genet.* **225,** 81–93.
26. Aragão, F. J. L., Ribeiro, S. G., Barros, L. M. G., et al. (1998) Transgenic beans (*Phaseolus vulgaris* L.) engineered to express viral antisense RNAs show delayed and attenuated symptoms to bean golden mosaic geminivirus. *Mol. Breed.* **4,** 491–499.
27. Singsit, C., Aadang, M. J., Lynch, R. E., et al. (1997) Expression of a *Bacillus thuringiensis* cryIA(c) gene in transgenic peanut plants and its efficacy against lesser cornstalk borer. *Transgenic Res.* **6,** 169–176.
28. Lius, S., Manshardt, R. M., Fitch, M. M. M., Slightom, J. L., Sanford, J. C., and Gonsalves, D. (1997) Pathogen-derived resistance provides papaya with effective protection against papaya ringspot virus. *Mol. Breed.* **3,** 161–168.
29. Bommineni, V. R., Chibbar, R. N., Bethune, T. D., Tsang, E. W. T., and Dunstan, D. I. (1997) The sensitivity of transgenic spruce (*Picea glauca* (Moench Voss) cotyledonary somatic embryos and somatic seedlings to kanamycin selection. *Transgenic Res.* **6,** 123–131.
30. Takenaka, M., Yamaoka, S., Hanajiri, T., et al. (2000) Direct transformation and plant regeneration of the haploid liverwort *Marchantia polymorpha* L. *Transgenic Res.* **9,** 179–185.

31. Koncz, C. and Schell, J. (1986) The promoter T_L-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Mol. Gen. Genet.* **204,** 383–396.
32. Cahill, M. A., Ernst, W. H., Janknecht, R., and Nordheim, A. (1994) Regulatory squelching. *FEBS Lett.* **344,** 105–108.
33. Lloyd, A. M., Walbot, V., and Davis, R. W. (1992) *Arabidopsis* and *Nicotiana anthocyanin* production activated by maize regulators, *R* and *C1*. *Science* **258,** 1773–1775.
34. Payne, C. T., Zhang, F., and Lloyd, A. M. (2000) GL3 encodes a bHLH protein that regulates trichome development in *Arabidopsis* through interaction with GL1 and TTG1. *Genetics* **156,** 1349–1362.
35. Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., and Lepiniec, L. (2000) The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in *Arabidopsis siliques*. *Plant Cell* **12,** 1863–1878.
36. Lee, M. M. and Schiefelbein, J. (2001) Developmentally distinct MYB genes encode functionally equivalent proteins in *Arabidopsis*. *Development* **128,** 1539–1546.
37. Sambrook, J. and Russell, D. W. (2001) *Molecular Cloning: A Laboratory Manual.* CSP Laboratory Press, Cold Spring Harbor, NY.
38. Jefferson, R. A., Kavanagh, T. A., and Bevan, M. W. (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J.* **6,** 3901–3907.
39. Jefferson, R. A. (1987) Assaying chimeric genes in plants: the GUS gene fusion system in plants. *Plant Mol. Biol. Rep.* **5,** 387–405.
40. Mitsuhara, I., Ugaki, M., Hirochika, H., et al. (1996) Efficient promoter cassettes for enhanced expression of foreign genes in dicotyledonous and monocotyledonous plants. *Plant Cell Physiol.* **37,** 49–59.
41. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S., and Mullineaux, P. M. (2000) pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **42,** 819–832.
42. Bevan, M. (1984) Binary vectors for plant transformation. *Nucleic Acids Res.* **12,** 8711–8721.
43. Höfgen, R. and Willmitzer, L. (1990) Biochemical and genetic analysis of different patatin isoforms expressed in various organs of potato. *Plant Sci.* **66,** 221–230.
44. An, G., Ebert, P., Mitra, A., and Ita, S. (1988) Binary vectors, in *Plant Molecular Biology Manual* (Gelvin, S. B. and Schilperoort, R. A., eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 1–19.
45. Schardl, C. L., Byrd, A. D., Benzion, G., Altschuler, M. A., Hildebrand, D. F., and Hunt, A. G. (1987) Design and construction of a versatile system for the expression of foreign genes in plants. *Gene* **61,** 1–11.

21

T-DNA Activation Tagging

Johan Memelink

Summary

T-DNA activation tagging is a method to generate dominant mutations in plants or plant cells by random insertion of a T-DNA carrying constitutive enhancer elements, which can cause transcriptional activation of flanking plant genes. The method consists of generating a large number of transformed plants or plant cells using a specialized T-DNA construct, followed by selection for the desired phenotype. Subsequently, the activated plant gene is rescued from selected mutant transformants for further functional analysis. Since the exact procedure depends on the plant material and the selected phenotype, this chapter describes one specific example of T-DNA activation tagging of suspension-cultured cells, including, where possible, cross-references to more general applications of the technique.

Key Words

agrobacterium, *Arabidopsis*, *Catharanthus roseus*, cauliflower mosaic virus 35S, cell suspension, 4-methyl tryptophan, plasmid rescue

1. Introduction

One of the most direct ways of dissecting complex biological processes in plants is generation and analysis of genetic mutants. Mutations, resulting from T-DNA or transposon tagging or chemical mutagenesis, usually cause loss of function and are, therefore, recessive. Consequently, the mutant phenotype can only be observed following selfing of the mutated plants. This demands a substantial amount of effort and is not possible for all plant species. Another drawback of loss-of-function mutagenesis is that mutation of functionally redundant genes does not lead to phenotypically altered plants.

Many of these disadvantages are circumvented by an alternative approach to generate mutants, called T-DNA activation tagging. A T-DNA carrying a

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

strong constitutive promoter reading towards one of its borders is introduced into plant cells. Upon random T-DNA integration into the genome, flanking plant sequences can be transcribed, which can result in a dominant mutation. Therefore, in contrast to classic mutagenesis, mutants generated by T-DNA activation tagging allow direct selection for the desired phenotype in the primary transformants. Furthermore, a phenotype can result from T-DNA activation tagging of a functionally redundant gene, allowing its analysis and cloning. Since it generates dominant mutations and introduces a DNA tag near the affected gene, activation tagging can be applied to genetically nontractable plant species for gene identification and isolation.

T-DNA activation tagging can be applied to plants as well as cultured cells. A disadvantage of plants is that it requires the generation of a large number of independent transformants (*see* **Note 1**). This is time-consuming even for *Arabidopsis thaliana*, and essentially impossible with other plant species. For laboratories that do not study *Arabidopsis* and do not have the manpower and growth facilities for large-scale plant transformation and regeneration, plant cell cultures are an alternative system for activation tagging. With cultured cells that are susceptible to transformation via *Agrobacterium tumefaciens*, the generation of large numbers of independent transformants is relatively easy. A clear disadvantage of cultured cells is that the number of screenable phenotypes is relatively small. One can, for example, screen for phytohormone-independent growth ***(1)*** or regeneration ***(2)*** or for the accumulation of visible metabolites.

The range of screenable phenotypes with cultured cells can be increased by using gene expression screens. In the example described here in detail, *Catharanthus roseus* (Madagascar periwinkle) cells were screened for resistance to toxic levels of the tryptophan derivative 4-methyl tryptophan (4mT) and, thus, for high expression levels of the 4mT-detoxifying enzyme tryptophan decarboxylase (TDC). This type of screen can probably be extended considerably by applying T-DNA activation tagging to transgenic cells containing a fusion between promoter sequences of interest and a selectable gene. In this way, gain-of-function mutations in signal transduction pathways activating the selected promoter can be identified.

Usually, T-DNA activation tagging selects for new phenotypes, which are due to overexpression of an activating component of a signal transduction pathway, such as a transcriptional activator. To search specifically for the phenotypes caused by overexpression of a repressor, the selection method can be adapted by using a negative selectable marker fused to a promoter, which is switched on by the signal transduction pathway of interest. Examples of negative selectable markers are the bacterial *cytosine deaminase* gene ***(3)*** and the T-DNA *tumor morphology shoots 2* gene ***(4)***. However, use of heterologous

positive or negative selection markers for plant activation tagging has not yet been described.

Since the transformation protocol depends on the plant material chosen and the selection method depends on the process studied and the desired phenotype, I will give a detailed description of the method, which was used in my group in an attempt to isolate regulators of terpenoid indole alkaloid (TIA) biosynthesis genes in *C. roseus*. This will give a good idea of the general approach, expected results, and the nature of possible control experiments.

1.1. Outline of the Experimental Procedure

For T-DNA activation tagging of *Arabidopsis* plants, ref. ***5*** gives a good description of the procedure. The method, described in detail here, concerns T-DNA activation tagging of cultured cells.

The gene encoding the TIA biosynthetic enzyme TDC was used as marker for mutant selection ***(6)***. TDC converts L-tryptophan into tryptamine, one of the first steps in TIA biosynthesis. TDC can use certain L-tryptophan derivatives as a substrate, such as 4mT ***(7)***. This compound is toxic for plant cells and is converted by TDC into the nontoxic 4-methyl-tryptamine. *C. roseus* suspension-cultured cells were transformed with a T-DNA construct carrying enhancer elements from the cauliflower mosaic virus (CaMV) 35S RNA promoter located near the right border and, subsequently, selected on 4mT. Resistant cell lines were further screened for high expression of the *TDC* gene and of a second coordinately regulated TIA biosynthesis gene *STR* (strictosidine synthase) by Northern blot analysis. This research strategy has resulted in the isolation of *ORCA3*, a gene encoding an AP2/ERF domain transcription factor, which regulates several genes involved in primary and alkaloid metabolism in *C. roseus* ***(8–10)***. The method described here for periwinkle cells was carried out according to the following protocol, and the number of lines that met the selection criterion in each subsequent step are in parentheses.

1. Selection of an appropriate plant cell line, *Agrobacterium* strain, and tagging vector.
2. Determination of the 4mT selection window (*see* **Note 2**).
3. *Agrobacterium*-mediated transformation of *C. roseus* cells (estimated number of stable transformants of 400,000–500,000).
4. Selection of 4mT-resistant calli (281 calli).
5. Conversion of calli to cell suspension lines (successful for 180 calli).
6. Screening of cell suspension lines for a high *TDC* gene expression level by Northern blotting (20 cell lines with high *TDC* expression).
7. Screening of *TDC*-expressing cell lines for a high *STR* gene expression level (six cell lines with high *TDC* and *STR* expression).
8. Isolation of chromosomal DNA from lines with desired phenotype.

9. Determination of T-DNA copy number by Southern blotting (four single-copy lines).
10. Isolation of flanking plant DNA from single-copy lines by plasmid rescue (successful for two lines).
11. Confirmation of the tag-induced phenotype by retransformation of plant cells by particle bombardment (successful for one line *[8]*).

Some theoretical and practical considerations for the choice of tagging vector, *Agrobacterium* strain, and plant cells are discussed below.

1.2. Choice of Plant Material

Activation tagging has been described for a number of plant species, including *Arabidopsis* ***(1,2,5,11–15)***, *Craterostigma plantagineum* ***(16)***, petunia ***(17)***, and *C. roseus* ***(8,10)***. With *Arabidopsis*, the method has also been used to tag and activate genes that suppress or modify a mutant phenotype ***(5,18–21)***. Different cell types have been used for *Agrobacterium*-mediated transformation, including protoplasts ***(17)***, leaf explants ***(12,16)***, root explants ***(2)***, calli ***(1)***, and suspension-cultured cells ***(8)***. For *Arabidopsis*, floral dip transformation ***(22,23)*** is a simple and efficient method to generate large numbers of transformed plants ***(5)***. This method has also potential for other plant species ***(23,24)***.

1.3. Choice of Agrobacterium Strain

The preferred transformation method for activation tagging is via *Agrobacterium*. It has the advantage that the majority of transformants have integrations of single copies of full-length T-DNA, which enormously simplifies further analysis of mutant lines.

Different *Agrobacterium* strains (**Table 1**) have been described. Details about *Agrobacterium* strains can be found in ref. ***25***. For activation tagging of *Arabidopsis*, strain GV3101::pMP90RK is often used. Certain plant species are more susceptible to certain *Agrobacterium* strains. Finding the most efficient strain for transformation of the selected plant material is a matter of trial and error. For *C. roseus* cell suspensions, a ternary *Agrobacterium* strain containing a constitutively active mutant version of *virG* turned out to be very effective ***(26)***.

1.4. Choice of Tagging Vector

The general structure of an activation tagging T-DNA is shown in **Fig. 1**. The most important feature is the presence of transcriptional enhancer elements or a promoter at the right border reading outward. An antibiotic resistance marker is present for selection of transformants. To enable plasmid rescue of

Table 1
Choice of *Agrobacterium* Strain[a]

Strain	Reference	Antibiotic markers[b]
GV3101::pMP90RK	***(5)***	rif, gent, and kan
LBA4404::pBBR*virG*N54D	***(8,10,26)***	rif, cam, spec, and strep
GV2260	***(12)***	rif, carb
ABI	***(2)***	cam, gent, and kan

[a]See ref. ***25*** for more information on some of these strains.
[b]Abbreviations of resistance markers: cam, chloramphenicol; carb, carbenicillin; gent, gentamycin; kan, kanamycin; rif, rifampicin; spec, spectinomycin; strep, streptinomycin.

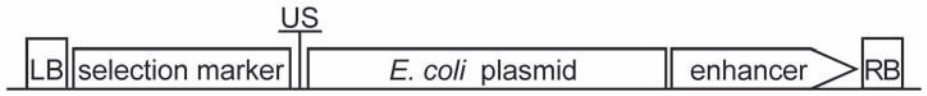


Fig.1. Schematic representation of an activation tagging T-DNA. LB, left border; RB, right border; US, unique restriction sites.

tagged genes, the T-DNA contains a set of unique restriction sites and a complete *Escherichia coli* plasmid (a pUC derivative or similar plasmid). Digestion of chromosomal DNA from a T-DNA-tagged mutant with one of the uniquely cutting restriction enzymes, followed by circularization, and transformation of *E. coli*, will rescue a portion of the flanking plant DNA.

Different tagging vectors differ mainly in the selectable marker for plant transformation (kanamycin, glufosinate, or hygromycin resistance) and the tagging promoter (**Table 2**). Three general types of tagging promoters can be envisaged, consisting of: *(i)* a complete promoter with TATA box and start codon; *(ii)* a complete promoter with TATA box without start codon; or *(iii)* a set of enhancer elements without TATA box. Each promoter has specific advantages. A tagging promoter with start codon can cause overexpression of truncated proteins, which may reveal phenotypes that are not easily detected with the complete protein. A TATA box-containing promoter can cause transcription of antisense RNA when integrated in reverse orientation downstream of a gene and, thereby, knock out gene function. However, most published tagging experiments have been performed with enhancer elements, because this gives the largest chance of finding phenotypes, since less stringent requirements are imposed on the integration site.

The enhancer elements used are often derived from the CaMV 35S promoter. They consist of four copies of the B domain of the CaMV 35S promoter ***(1,5)***, two copies of the B domain fused to four copies of the A1 domain ***(8,27)***,

Table 2
Choice of Tagging Vectors

Plasmid	Reference	Bacterial marker[a]	Plant marker[a]	Enhancer[b]	TATA box
pPCVICEn4HPT	***(1)***	carb	hyg	4B	–
Tag2B4A1	***(8,10)***	kan	hyg	2B + 4 A1	–
pSKI015	***(5)***	carb	bar	4B	–
pSKI074	***(5)***	carb	kan	4B	–
pSDM1550	***(12)***	kan	hyg	2B + A	+
pER16	***(2)***	spec	kan	estradiol-inducible	+
pCVHPT	***(16)***	carb	hyg	gene 5 enhancer	–
pMON29963	***(14)***	spec/strep	kan	4B	–

[a]Abbreviations of resistance markers: bar, bialaphos (or phosphinothricin or glufosinate); carb, carbenicillin; hyg, hygromycin; gent, gentamycin; kan, kanamycin; spec, spectinomycin; strep, streptinomycin.

[b]Enhancers consist of single or multiple copies of the B, A1, or A domains of the CaMV 35S promoter, the T-DNA gene 5 enhancer, or an estradiol-inducible promoter system.

or two copies of the B domain fused to one copy of the A domain including the TATA box ***(12)***. Other enhancer elements described in the literature include the T-DNA gene 5 enhancer ***(16)***. Most of the published activation tagging experiments with *Arabidopsis* use the vectors with four copies of the B domain described by ref. ***5*** (*see* **Note 3**). In addition, a chemical-inducible activation tagging system has been used in *Arabidopsis* ***(2)***, which has the advantage that withdrawal of the inducer β-estradiol allows recovery of morphologically normal mutant plants in those events in which the gain-of-function mutation causes developmental defects or lethality.

2. Materials

2.1. Choice of Agrobacterium *Strain*

Agrobacterium strains which have been used in published tagging experiments include the ones listed in **Table 1**. Other strains described in ref. ***25*** might also be suitable in combination with certain tagging vectors (*see* **Table 2**).

2.2. Choice of Tagging Vector

Various tagging vectors described in the literature are listed in **Table 2**.

2.3. Catharanthus *Cell Cultures*

1. *Catharanthus* cell lines are grown in 20–25 or 55–75 mL vol in 100- or 250-mL Erlenmeyer flasks, respectively, with foam stoppers in a 16:8 h light:dark cycle at 25°C on a gyratory shaker at 125 rpm.
2. Untransformed cell line BIX is subcultured weekly by transferring 25mL of cells to 50 mL of LS-13 medium in an 250-mL Erlenmeyer flask.
3. Subculturing of transgenic BIX lines is done weekly by 7.5-fold dilution in LS-13 medium containing 50 μg/mL hygromycin. For lines transformed with *Agrobacterium*, the medium also contains 400 μg/mL cefotaxime and 100 μg/mL vancomycin for the first 10 subculture cycles or as long as needed.

2.4. *Preparation of* Agrobacterium *Culture*

1. AB medium: to prepare 1 L of AB medium ***(28)***, autoclave 900 mL water containing 5 g glucose at 110°C for 30 min. Solid medium contains in addition 16 g agar/L. After autoclaving, add 50 mL 20× AB salts and 50 mL 20× AB buffer.
   a. 20× AB buffer: 60 g K_2HPO_4, 20 g NaH_2PO_4/L, autoclave at 120°C.
   b. 20× AB salts: 20 g NH_4Cl, 6 g $MgSO_4 \cdot 7H_2O$, 3 g KCl, 0.2 g $CaCl_2$, 50 mg $FeSO_4 \cdot 7H_2O$/L, autoclave at 120°C.
2. Rifampicin (Duchefa): 20 mg/mL in methanol. Store at –20°C.
3. Kanamycin (Duchefa): 100 mg/mL in water. Filter-sterilize through a 0.22-μm membrane (Millipore, Bedford, MA, USA) and store at –20°C.
4. Chloramphenicol (Duchefa): 75 mg/mL in ethanol. Store at –20°C.

2.5. *Co-Cultivation*

1. LS-13 medium ***(29)***: dissolve the prescribed amount of Linsmaier-Skoog medium powder (Duchefa) and 30 g sucrose/L. The solution will contain 2.99 m*M* $CaCl_2$, 1.25 m*M* KH_2PO_4, 18.79 m*M* KNO_3, 1.50 m*M* $MgSO_4$, 20.61 m*M* NH_4NO_3, 0.11 μ*M* $CoCl_2$, 0.10 μ*M* $CuSO_4$, 0.10 m*M* FeNaEDTA, 0.10 m*M* H_3BO_3, 5.00 μ*M* KI, 0.10 m*M* $MnSO_4$, 1.03 μ*M* Na_2MoO_4, 29.91 μ*M* $ZnSO_4$, 0.56 m*M* myoinositol, 1.19 μ*M* thiamine-HCl. Add 2 mg/L 1-NAA and 0.2 mg/L kinetin from liquid 100× stocks. pH is adjusted to 5.8 with KOH. Solid medium contains 0.7% plant tissue culture agar (Imperial laboratories) or Daishin agar (Brunschwig Chemie). Store sterile liquid or solidified medium at 4°C.
2. Co-cultivation medium: LS-13 medium supplemented with 10 g/L glucose. pH is adjusted to 5.2. Solid medium contains 0.7% plant tissue culture agar. Autoclave at 110°C for 30 min. Store at 4°C. Acetosyringone (*see* **Note 4**) is added after autoclaving to a final concentration of 100 μ*M*.
3. Hygromycin selection medium: LS-13 medium pH 5.8. Solid medium contains 0.7% plant tissue culture agar. After sterilization, 400 mg/L cefotaxime, 100 mg/L vancomycin, and 50 mg/L hygromycin are added from 1000× stocks.
4. 4mT Selection medium: hygromycin selection medium containing 4mT. 4mT (Sigma) is added as a powder prior to autoclaving. Solid 4mT selection medium contains 0.4 mM 4mT, whereas liquid 4mT selection medium contains 0.2 mM 4mT (*see* **Note 2**).

5. Kinetin (Research Organics) and 1-NAA (BDH Chemicals): 100× stocks are 20 and 200 mg/L, respectively. The hormones are dissolved initially in a small vol of ethanol, water is added, and the ethanol is removed by boiling. Stocks can be kept at 4°C for several mo.
6. Acetosyringone (3',5'-dimethoxy-4'-hydroxy-acetophenone): 100 m*M* in dimethyl sulfoxide (DMSO) (both from Sigma). Store at –20°C.
7. Hygromycin (Calbiochem-Novabiochem.): 50 mg/mL in water. pH is adjusted to 7.0 with 1 *M* HCl. The filter-sterilized solution can be kept for several mo at 4°C.
8. Cefotaxime (Duchefa): 400 mg/mL in water. The filter-sterilized solution can be kept for several mo at 4°C.
9. Vancomycin (Duchefa): 100 mg/mL in water. The filter-sterilized solution can be kept for several mo at 4°C.

2.6. Plasmid Rescue

1. Phenol: melt analytical quality solid phenol (Merck) at 60°C. Add 1 vol of 0.1 *M* Tris-HCl, pH 8.0, and 1 g 8-hydroxyquinoline (Merck)/L liquid phenol. This antioxidant will dissolve in the phenol phase giving it a yellow color. Mix the two phases by shaking vigorously. Wait until the two phases are completely separated, remove most of the aqueous phase, and add another vol of 0.1 *M* Tris-HCl, pH 8.0. Shake vigorously. The solution can be stored for several mo at 4°C in the dark.
2. 3 *M* Sodium acetate, pH 4.8: the pH is adjusted with acetic acid.
3. 10× ligation buffer: 500 m*M* Tris-HCl, pH 7.5, 10 m*M* ATP, 100 m*M* $MgCl_2$, 100 mM dithiothreitol (DTT).
4. $T_{10}E_1$: 10 m*M* Tris-HCl, pH 7.5, 1 m*M* ethylene diamine tetraacetic acid (EDTA).
5. *E. coli* strain NM554 [F^- *araD139 Δ(ara-leu)7696 galE15 galK16 Δ(lac)X74 rpsL* (Str^r) *hsdR2* (r_K^- m_K^+) *mcrA mcrB1 recA13*] or other suitable *E. coli* host strain, which is deficient in the restriction systems hsdR, mcrA, and mcrCB and is *recA*$^-$.

2.7. Electroporation of E. coli Cells

1. LC medium: 10 g tryptone (Difco), 5 g yeast extract (Difco), 8 g NaCl/L. Autoclave for 20 min at 120°C. Solid medium contains 1.6% agar.
2. SOC medium: 2% tryptone, 0.5% yeast extract, 10 m*M* NaCl, 2.5 m*M* KCl, 10 m*M* $MgCl_2$, 10 m*M* $MgSO_4$, 20 m*M* glucose. Autoclave for 20 min at 120°C.
3. Carbenicillin (Duchefa): 200 mg/mL in water. Filter-sterilize and store at –20°C.

3. Methods

3.1. Agrobacterium-Mediated Transformation of Suspension-Cultured Cells

Handling of suspension-cultured plant cells and of bacterial strains is carried out under sterile conditions using autoclaved or filter-sterilized solutions, autoclaved glassware, and sterile disposables.

Susceptibility and efficiency of the selected plant material for *Agrobacterium*-mediated transformation is tested first (*see* **Notes 5–7**). In addition, the enhancer elements used for tagging can be tested for their transcriptional activity in the selected plant material (*see* **Note 8**).

3.1.1. Preparation of Agrobacterium Culture

1. Inoculate *Agrobacterium* strain LBA4404::pBBR*virG*N54D::Tag2B4A1 on solid AB medium supplemented with 20 μg/mL rifampicin, 100 μg/mL kanamycin, and 75 μg/mL chloramphenicol, and grow for 3 d at 29°C.
2. Use several *Agrobacterium* colonies from this plate to inoculate 10 mL of liquid AB medium containing the same antibiotics and grow overnight on a shaker at 29°C.
3. Measure the OD_{600} of this culture and transfer it to a sterile tube.
4. Centrifuge at 3000*g* for 5 min and remove the supernatant immediately.
5. Resuspend the bacterial pellet in co-cultivation medium to an OD_{600} of 1.

3.1.2. Preparation of C. roseus Cells

1. Transfer 7-d-old cells of cell line BIX to 50-mL conical plastic tubes and let the cells settle for 5–10 min.
2. Replace the medium by an equal vol of co-cultivation medium.
3. Repeat this medium replacement 2×.
4. Adjust the final vol to settled cell vol of 50%, corresponding to a cell density of 1 to 2×10^6 cells/mL.

3.1.3. Co-Cultivation

1. Prepare Petri dishes (94 × 25 mm) with solid co-cultivation medium. Leave the plates open to dry for about 1 h.
2. Mix on the plates 7 mL of *C. roseus* cells with 700 μL *Agrobacterium* suspension.
3. Seal the Petri dishes with urgopore tape (Chenove) and incubate them in the dark at 25°C for 3–5 d (*see* **Note 9**).
4. After 3–5 d, almost all the liquid will be absorbed by the medium. Add 25 mL of LS-13, mix, and collect the cells on Whatman filters (90-mm diameter, Whatman No. 5; Whatman) in a Buchner funnel.
5. Wash the cells once more with 25 mL LS-13 to remove the majority of the Agrobacteria.
6. Place the filters carrying the cells in a thin even layer on Petri dishes (94 × 16 mm) with solid 4mT selection medium and incubate at 25°C in a 16:8 light:dark cycle until calli appear (takes 5–10 wk).
7. To get an indication of transformation efficiency, one filter is placed on hygromycin selection medium. A more precise estimation of transformation efficiency can be obtained as outlined in **Note 7**.
8. To confirm the selectivity of the procedure, *C. roseus* cells are transformed with

Agrobacterium strain LBA4404::pBBR*virG*N54D carrying a binary plasmid with the hygromycin selectable marker, but lacking the tagging enhancer, and plated on solid 4mT selection medium. No or very few calli should appear. These (false positive) calli usually do not grow upon transfer to solid 4mT selection medium. To verify that the transformation procedure was successful, cells from this control co-cultivation are also plated on solid hygromycin selection medium.

3.2. Subculturing of Resistant Calli

3.2.1. Subculturing of Calli

1. Transfer resistant calli on solid 4mT selection medium and grow for 2–4 wk.
2. Compare growth with negative and positive control calli (*see* **Note 10**).
3. Subculture resistant calli once more before molecular analysis.
4. Calli can now be analyzed directly (*see* **Notes 11** and **12**) or be converted to cell suspensions and then analyzed.

3.2.2. Conversion of Calli into Cell Suspensions

1. Grow calli to a diameter of 1.5 cm. Transfer a small amount of each (numbered) callus to a separate plate as a backup. Transfer backups every 3 wk.
2. Transfer the rest of the callus in 5 mL of liquid 4mT selection medium in a 100-mL Erlenmeyer flask. Disperse cell clumps with a forceps. This will immediately create a cell suspension-like mixture.
3. Incubate on a shaker. Add 5 mL fresh 4mT selection medium when cell density becomes high. Gradually increase the culture vol to 20 mL (*see* **Note 13**).
4. When a vol of 20 mL of dense suspension is reached (after about 2 wk), 3 mL of cells are transferred weekly to 20 mL of 4mT selection medium.

3.3. Plasmid Rescue from Tagged Lines

3.3.1. DNA Isolation

1. DNA is isolated using a procedure yielding high molecular weight DNA, which is digestible with restriction enzymes (*see* **Note 14**).
2. Ten micrograms of DNA is digested with different restriction enzymes selected from the set of uniquely cutting enzymes (**Fig. 1**) and analyzed by Southern blot for T-DNA copy number. The Southern blot also predicts the expected size of flanking plant DNA, which can be isolated by plasmid rescue (*see* **Note 15**).

3.3.2. Plasmid Rescue

1. Five micrograms of chromosomal DNA from *C. roseus* is digested overnight with 100 U of restriction enzyme according to the manufacturer's instructions in a 250 µL vol (*see* **Note 15**).
2. The restriction mixture is extracted once with 1 vol of a 1:1 mixture of phenol/$CHCl_3$ and once with $CHCl_3$.
3. DNA is precipitated by addition of 0.1 vol of 3 *M* sodium acetate, pH 4.8, and 2 vol of ethanol, washed with 70% ethanol, dried, and resuspended in 20 µL $T_{10}E_1$.

4. One microliter is run on a 1% agarose/1× Tris-borate EDTA (TBE) gel to check the DNA amount and the degree of digestion in comparison with undigested DNA.
5. Ligation is carried out overnight at 14°C at low DNA concentration in a vol of 1 mL by addition of 100 µL 10× ligation buffer, water, and 2 U of T4 DNA ligase.
6. Extract the ligation mixture once with 1 vol of a 1:1 mixture of phenol/$CHCl_3$ and once with $CHCl_3$.
7. DNA is precipitated by addition of 0.1 vol of 3 *M* sodium acetate, pH 4.8, and 2 vol of ethanol, washed with 70% ethanol, dried, and resuspended in 20 µL of water.
8. One microliter is run on an agarose gel to check the DNA amount and the formation of high molecular weight ligation products.
9. Since it is very easy to pick up plasmid contaminations, a control digestion is carried out on 5 µg of salmon sperm DNA, following exactly the same procedure. This control checks whether the enzymes, corresponding buffers, and other solutions used are plasmid-free (*see* **Note 16**).
10. Another control is direct transformation of 5 µg of undigested *C. roseus* DNA. This control checks whether the isolated DNA is plasmid-free.
11. If the negative controls yield transformants, adequate measures should be taken to avoid plasmid contamination (*see* **Note 16**).

3.3.3. Preparation of Electrocompetent E. coli Cells

1. Dilute an overnight culture of *E. coli* strain NM554 100-fold in 100 mL LC medium.
2. Grow to an OD_{600} of about 0.5 and place on ice for 15 min.
3. Centrifuge for 15 min at 1000*g* at 4°C.
4. Wash 2× with 50 mL of ice-cold water.
5. Wash once with 5 mL of ice-cold 10% glycerol.
6. Resuspend cells in 400 µL of 10% glycerol/10% polyethylene glycol (PEG)4000.
7. Freeze 40-µL aliquots in liquid N_2 and store at –80°C.
8. Check competence with 1 ng of pBluescript® plasmid (Stratagene) (*see* **Note 17**).

3.3.4. Electroporation of E. coli Cells

1. Thaw cells on ice and chill new electroporation cuvettes with a 2-mm electrode gap on ice.
2. Mix 40 µL of cells with 5 µL of ligated DNA on ice, including the appropriate negative controls.
3. Transfer the cells to an ice-cold electroporation cuvette. Suspension should contact both electrodes.
4. Apply a 4-ms electric pulse at settings of 2.5 kV, 200 Ohm, and 25 µF (*see* **Note 17**).
5. Immediately add 1 mL of SOC medium and gently mix the cells.
6. Transfer to a 1.5-mL tube and incubate for 1 h at 37°C.
7. Collect cells by centrifugation at 1000*g* for 1 min and resuspend in 100 µL of LC medium.

8. Plate on Petri dishes with solid LC medium containing 200 μg/mL carbenicillin (*see* **Note 18**).
9. Plasmid DNA from antibiotic-resistant colonies can be analyzed by restriction enzyme digestion (*see* **Note 16**), polymerase chain reaction (PCR), and sequencing.
10. The ability of the rescued flanking plant DNA to confer the desired phenotype is tested by transforming plant cells (*see* **Note 19**).

4. Notes

1. To obtain a good idea of the number and type of mutants that one can obtain via T-DNA activation tagging, a near saturating screen is necessary. The number of independent transformants that needs to be screened to obtain a high probability of tagging any given gene depends on genome size and can be estimated using the formula: $n = [\log_{10} (1–P)]/[\log_{10} (1–[x/\text{genome size}])]$; where P = probability value between 0 and 1, x = length of the targeted DNA region in kb, n = number of T-DNA insertions in the population, and haploid genome size is in kb. The formula assumes that T-DNA integration is random. Activation tagging mutants have been found with the tag inserted at a distance of 3.6 kb from the overexpressed gene ***(5)***. If we apply the formula to *Arabidopsis* with a haploid genome size of 1.25×10^8 bp, assuming that T-DNA integration within a 3-kb region should activate a gene and that transformants contain on average 2 T-DNA copies and if we are satisfied with a 90% probability of tagging of any given gene, the result is that approx 50,000 transformants need to be screened.
2. To establish the effective 4mT concentration for selection of *C. roseus* cell lines with increased TDC enzyme activity, the 4mT selection window was determined. *C. roseus* cell suspensions were transformed with the *TDC* gene under control of the CaMV 35S promoter and, as a negative control, with an empty vector carrying only the hygromycin resistance gene under conditions mimicking the transformation conditions in the actual tagging experiment. We reasoned that the CaMV 35S-driven expression of the *TDC* gene would be the highest expression level we would be able to obtain via random tagging. Transformation mixtures were plated on hygromycin selection medium and a range of 4mT concentrations. Cells overexpressing the *TDC* gene showed normal growth on concentrations up to 0.4 m*M* 4mT, whereas growth of control cells was retarded at concentrations of 0.1 and 0.2 m*M* and completely inhibited at 0.4 m*M* ***(10)***. Based on these experiments, 0.4 m*M* 4mT was chosen as the selective condition.
3. Progressive loss of enhancer copies from the 4B-type vectors due to homologous recombination has been described upon storage of *Agrobacterium* at 4°C ***(5)***. It is recommended to start bacterial cultures for plant cell transformation from a fresh inoculum taken from a –80°C stock and to verify the presence of all four B enhancer domains by PCR. We have never observed loss of enhancer copies from the 2B4A1-type tagging vector.
4. Acetosyringone is added in all co-cultivations, even though certain *Agrobacterium* strains (such as LBA4404::pBBR*virG*N54D) contain constitutively active *virG*

versions. If acetosyringone has a negative effect on the suspension-cultured cells, leave it out and switch to acetosyringone-independent strains.

5. To test the susceptibility and efficiency of the plant material to transformation with the selected *Agrobacterium* strain, the strain is provided with a binary plasmid carrying an intron-containing β-glucuronidase (GUS) or green fluorescent protein (GFP) reporter gene controlled by the CaMV 35S promoter. Reporter gene activity is visualized after co-cultivation. The presence of the intron ensures that reporter gene activity is due to gene expression *in planta*, and, thus, forms a marker for transient T-DNA transfer. Although only a small portion of the initially transferred T-DNA molecules is stably integrated, transient T-DNA transfer frequency usually correlates well with stable transformation efficiency.
6. A number of *C. roseus* cell lines were hypersensitive to *Agrobacterium* strain LBA4404 and died. Cell line BIX, on the other hand, tolerated exposure to *Agrobacterium* without apparent stress symptoms. Therefore, it is worthwhile to test different cell lines of a plant species for hypersensitivity to *Agrobacterium*. It is possible that hypersensitivity also depends on the *Agrobacterium* strain, although we did not test this.
7. Stable transformation frequencies can be estimated by plating cells on hygromycin selection medium. For estimation of transformation frequencies of cell suspension cells, co-cultivated cells need to be diluted in a logarithmic series with untransformed cells. Stable transformation frequencies can be as high as 1500 calli/mL of initial cells, corresponding to 10,000 calli/94-mm Petri dish ***(26)***.
8. We tested the activity of the 2B4A1 enhancer, which is present on the tagging vector Tag2B4A1, by fusing it to the GUS reporter gene in the vector GusXX-47 ***(30)***. The activity of the enhancer was compared with the activity of the –940 or the double-enhanced CaMV 35S promoter, in transient expression assays using particle bombardment and in stable transformants, and was found to be similarly active in *C. roseus* cells ***(27)***. The cell-specific activity of the 4B enhancer used in most published tagging experiments has never been described. However, a single copy of the B domain is active mainly in leaf tissue, with weaker activity in stem tissue and little activity in roots ***(31)***. A combination of the B and the A domain is highly active in most tissues. In activation-tagged *Arabidopsis* mutants, it was also reported that the 4B enhancer sometimes increased the normal gene expression level without changing tissue specificity, instead of giving constitutive expression ***(5)***. Therefore, depending on the plant tissue selected as a target for T-DNA activation tagging, it may be worthwhile to check enhancer activity.
9. Standard co-cultivation time is 3 d. With the combination of plant cells and *Agrobacterium* strain used here, no transient T-DNA expression was measurable at earlier time points. Sometimes, a co-cultivation time of 5 d gave better transformation results.
10. Negative control callus tissue can be taken from the hygromycin selection plate used to check transformation efficiency with the Tag2B4A1 vector (*see* **Subheading 3.1.3.**, **step 7**) or from the hygromycin selection plate used to check transformation efficiency with the negative control vector (*see* **Subheading 3.1.3.**, **step**

8). Positive control callus is generated by transformation of *C. roseus* cells with LBA4404::pBBR*virG*N54D carrying a binary plasmid containing the *TDC* gene under control of the CaMV 35S promoter ***(10)***.

11. Calli can be analyzed by Southern blot ***(32)*** or PCR analysis for T-DNA presence and copy number and/or by Northern blot ***(32)*** or reverse transcription PCR (RT-PCR) analysis for expression of genes of interest. DNA can be extracted from calli for isolation of T-DNA flanking plant sequences by thermal asymmetric interlaced PCR (TAIL-PCR) ***(33)*** or plasmid rescue. In the case of T-DNA activation tagging of *C. roseus* with 4mT selection, calli were converted into cell suspensions, which has the advantage that more biomass is rapidly obtained. With tagging vector Tag2B4A1, it is impossible to design TAIL-PCR primers close to the right border, due to the fact that it is flanked immediately by the B and A1 repeats. Since the 2B4A1 enhancer is 800 bp in size, we decided to use plasmid rescue. In general, rescue of flanking plant DNA by TAIL-PCR is difficult, because of the enhancer repeats and due to the fact that T-DNA transfer at the right border is often incomplete ***(5)***.
12. We screened the 4mT-resistant cell lines for high expression levels of the *TDC* gene, because we were interested in tagging of transcriptional regulators. In addition, we screened the cell lines for high expression levels of the *STR* gene, another TIA biosynthesis gene, which is coordinately regulated with the *TDC* gene. We reasoned that this last screen would select cell lines with a T-DNA tag flanking a transcriptional regulator of these two and possibly other TIA biosynthesis genes. These secondary screens reduced the number of cell lines for further analysis considerably. Out of 180 4mT-resistant cell lines, 20 expressed *TDC* at a high level, of which 6 also showed high *STR* expression ***(10)***. Therefore, it is worthwhile to design secondary screens, if possible, to select the most promising lines.
13. A relatively high cell density is important for rapid cell division. Therefore, at this stage, each cell suspension should be evaluated separately for dilution rate.
14. Chromosomal DNA purified by classic phenol extraction followed by CsCl-EtBrd centrifugation gave us excellent results, but requires a relatively large amount of tissue (which is not limiting with cell suspensions). We obtained no plasmid rescue transformants with DNA prepared with the Nucleon Phytopure DNA extraction kit (Amersham Pharmacia Biotech). However, in other people's hands and/or with other plant tissue, this kit or comparable kits, like the DNeasy plant kit (Qiagen) may give satisfactory results.
15. DNA amounts used for Southern blotting and plasmid rescue depend on the genome size of the plant species. As a rule of thumb for diploid cells, use about 0.5 μg for each 10^8 bp of haploid genome size for Southern blotting. A range of DNA amounts around this value can be tested for plasmid rescue efficiency. *C. roseus* is a diploid plant species with an estimated haploid genome size of about 10^9 bp. However, the BIX cell line was octaploid at the time of the tagging experiments. This may explain why it was difficult to obtain detectable signals in Southern blot hybridizations and may also explain the low number of rescued plasmids (*see* **Note 18**). All *C. roseus* cell lines tested had varying ploidy levels of 4*n* or more. Therefore, polyploidy seems to be a common feature of cell lines.

16. Try to avoid isolation of plasmid contaminations during all steps of the procedure, including plant DNA isolation, by using dedicated solutions and disposable material where possible. Use new enzymes and buffers, because normally, all enzymes and buffers in a common user stock are contaminated with plasmid DNA. In addition, it is convenient if the initial restriction enzyme analysis yields bands that are diagnostic for the rescued tagging vector. Rescued Tag2B4A1 plasmids, for example, give an 800-bp *Eco*RI band corresponding to the 2B4A1 enhancer.
17. Cells with adequate competence should give at least 5×10^8 transformants/μg of pBluescript plasmid. Electroporation settings can be experimentally optimized using 1 ng of pBluescript plasmid DNA.
18. Using cells with a competence of 5×10^8 transformants/μg of pBluescript plasmid, we obtained between 5–25 colonies in rescue experiments with different tagged cell lines and different restriction enzymes (*see* **Note 15**).
19. Plant cells can be directly transformed with the rescued plasmid via particle gun bombardment as described previously ***(27)***. Particles are coated with a mixture of the rescued plasmid and a plasmid containing a hygromycin resistance gene, e.g., pGL2 ***(34)***, in a 4 to 1 ratio. Alternatively, the flanking DNA can be subcloned from the rescued plasmid into a binary plant expression vector and introduced in plant cells via *Agrobacterium*-mediated transformation. Reference ***25*** provides some examples of binary plant expression vectors containing the CaMV 35S promoter, and ref. ***5*** describes two expression vectors containing the CaMV 35S B domain tetramer designed for recapitulating the mutant phenotype conferred by the flanking plant DNA. Controls include transformation with the empty expression vector.

References

1. Kakimoto, T. (1996) CKI1, a histidine kinase homolog implicated in cytokinin signal transduction. *Science* **274,** 982–985.
2. Zuo, J., Niu, Q.-W., Frugis, G., and Chua, N.-H. (2002) The *WUSCHEL* gene promotes vegetative-to-embryonic transition in *Arabidopsis*. *Plant J*. **30,** 349–359.
3. Schlaman, H. R. M. and Hooykaas, P. J. J. (1997) Effectiveness of the bacterial *codA* encoding cytosine deaminase as a negative selectable marker in *Agrobacterium*-mediated plant transformation. *Plant J.* **11,** 1377–1385.
4. Karlin-Neumann, G. A., Brusslan, J. A., and Tobin, E. M. (1991) Phytochrome control of the *tms2* gene in transgenic *Arabidopsis*: a strategy for selecting mutants in the signal transduction pathway. *Plant Cell* **3,** 573–582.
5. Weigel, D., Hoon Ahn, J., Blázquez, M. A., et al. (2000) Activation tagging in *Arabidopsis*. *Plant Physiol.* **122,** 1003–1013.
6. Goddijn, O. J. M., van der Duyn-Schouten, P. M., Schilperoort, R. A., and Hoge, J. H. C. (1993) A chimaeric tryptophan decarboxylase gene as a novel selectable marker in plants. *Plant Mol. Biol.* **22,** 907–912.
7. Sasse, F., Buchholz, M., and Berlin, J. (1983) Site of action of growth inhibitory tryptophan analogues in *Catharanthus roseus* cell suspension cultures. *Z. Naturforsch*. **38c**, 910–915.

8. van der Fits, L. and Memelink, J. (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science* **289,** 295–297.
9. van der Fits, L. and Memelink, J. (2001) The jasmonate-inducible AP2/ERF-domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate-responsive promoter element. *Plant J.* **25,** 43–53.
10. van der Fits, L., Hilliou, F., and Memelink, J. (2001) T-DNA activation tagging as a tool to isolate regulators of a metabolic pathway from a genetically non-tractable plant species. *Transgenic Res.* **10,** 513–521.
11. Kardailsky, I., Shukla, V., Ahn, J. H., et al. (1999) Activation tagging of the floral inducer *FT*. *Science* **286,** 1962–1965.
12. van der Graaff, E., den Dulk-Ras, A., Hooykaas, P. J. J., and Keller, B. (2000) Activation tagging of the *LEAFY PETIOLE* gene affects leaf petiole development in *Arabidopsis thaliana*. *Development* **127,** 4971–4980.
13. Borevitz, J. O., Xia, Y., Blount, J., Dixon, R. A., and Lamb, C. (2000) Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **12,** 2383–2393.
14. Huang, S., Cerny, R. E., Bhat, D. S., and Brown, S. M. (2001) Cloning of an *Arabidopsis* patatin-like gene, *STURDY*, by activation T-DNA tagging. *Plant Physiol.* **125,** 573–584.
15. Zhao, Y., Christensen, S. K., Fankhauser, C., et al. (2001) A role for flavin monooxygenase-like enzymes in auxin biosynthesis. *Science* **291,** 306–309.
16. Furini, A., Koncz, C., Salamini, F., and Bartels, D. (1997) High level transcription of a member of a repeated gene family confers dehydration tolerance to callus tissue of *Craterostigma plantagineum*. *EMBO J.* **16,** 3599–3608.
17. Zubko, E., Adams, C. J., Macháèková, I., Malbeck, J., Scollan, C., and Meyer, P. (2002) Activation tagging identifies a gene from *Petunia hybrida* responsible for the production of active cytokinins in plants. *Plant J.* **29,** 797–808.
18. Neff, M. M., Nguyen, S. M., Malancharuvil, E. J., et al. (1999) *BAS1*: a gene regulating brassinosteroid levels and light responsiveness in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **96,** 15316–15323.
19. Ito, T. and Meyerowitz, E. M. (2000) Overexpression of a gene encoding a cytochrome P450, *CYP78A9*, induces large and seedless fruit in *Arabidopsis*. *Plant Cell* **12,** 1541–1550.
20. Lee, H., Suh, S.-S., Park, E., et al. (2000) The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. *Genes Dev.* **14,** 2366–2376.
21. Li, J., Lease, K. A., Tax, F. E., and Walker, J. C. (2001) BRS1, a serine carboxypeptidase, regulates BRI1 signaling in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **98,** 5916–5921.
22. Clough, S. J. and Bent, A. F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16,** 735–743.
23. Bent, A. F. (2000) *Arabidopsis* in planta transformation. Uses, mechanisms, and prospects for transformation of other species. *Plant Physiol*. **124,** 1540–1547.

24. Trieu, A. T., Burleigh, S. H., Kardailsky, I. V., et al. (2000) Transformation of *Medicago truncatula* via infiltration of seedlings or flowering plants with *Agrobacterium*. *Plant J.* **22,** 531–541.
25. Hellens, R., Mullineaux, P., and Klee, H. (2000) A guide to *Agrobacterium* binary Ti vectors. *Trends Plant Sci*. **5,** 446–451.
26. van der Fits, L., Deakin, E. A., Hoge, J. H. C., and Memelink, J. (2000) The ternary transformation system: constitutive *virG* on a compatible plasmid dramatically increases *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **43,** 495–502.
27. van der Fits, L. and Memelink, J. (1997) Comparison of the activities of CaMV 35S and FMV 34S promoter derivatives in *Catharanthus roseus* cells transiently and stably transformed by particle bombardment. *Plant Mol. Biol*. **33,** 943–946.
28. Chilton, M.-D., Currier, T. C., Farrand, S. K., Bendich, A. J., Gordon, M. P., and Nester, E. W. (1974) *Agrobacterium tumefaciens* DNA and PS8 bacteriophage DNA not detected in crown gall tumors. *Proc. Natl. Acad. Sci. USA* **71,** 3672–3676.
29. Linsmaier, E. M. and Skoog, F. (1965) Organic growth factor requirements of tobacco tissue cultures. *Physiol. Plant.* **18,** 100–127.
30. Pasquali, G., Ouwerkerk, P. B. F., and Memelink, J. (1994) Versatile transformation vectors to assay the promoter activity of DNA elements in plants. *Gene* **149,** 373–374.
31. Benfey, P. N. and Chua, N.-H. (1990) The cauliflower mosic virus 35S promoter: combinatorial regulation of transcription in plants. *Science* **250,** 959–966.
32. Memelink, J., Swords, K. M. M., Staehelin, L. A., and Hoge, J. H. C. (1994) Southern, Northern and Western blot analysis, in *Plant Molecular Biology Manual* (Gelvin, S. B. and Schilperoort, R. A., eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. F1–F23.
33. Liu, Y.-G., Mitsukawa, N., Oosumi, T., and Whittier, R. F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J*. **8,** 457–463.
34. Bilang, R., Iida, S., Peterhans, A., Potrykus, I., and Paszkowski, J. (1991) The 3'-terminal region of the hygromycin-B-resistance gene is important for its activity in *Escherichia coli* and *Nicotiana tabacum*. *Gene* **100,** 247–250.

22

Expression Profiling Using cDNA Microarrays

Suling Zhao and Wesley B. Bruce

Summary

Microarray technology has become increasingly useful in measuring expression levels of a large number of genes and part of a repertoire of functional genomic tools. We describe the methods of cDNA microarray preparation, the use, data collection, and initial data processing. The cDNA fragments are first prepared by polymerase chain reaction (PCR), and then attached to a solid substrate, such as a chemically treated glass slide. Robotic machines spot the prepared cloned cDNA samples in a miniaturized gridded pattern, so that nanoliter amounts of tens of thousands cDNA samples are bound to a single 7.5 × 2.5 cm glass slide. Probes are generated from RNA samples of test and control tissues by incorporating Cyanine dyes (Cy™3 or Cy5) in reverse-transcribed products. Probes from a test sample are labeled with one of two Cy dyes and mixed in equal amounts with probes from a control sample labeled with the second dye. The glass slides containing the cDNA microarray are hybridized with the mixed Cy-labeled probes, washed, dried, and scanned using laser scanners with an optimized wavelength to excite each Cy dye. The emission image patterns for each dye are captured by a digital camera using micro-optics and processed into numerical values that positively correlate with quantitative levels of mRNA for each cDNA spot on the slide. The collected data is then further processed, normalized across experiments, and examined via numerous statistical and mathematical approaches to infer changes in expression levels of particular genes due to the treatment tested.

Key Words

cDNA microarray, gene expression profiling, Cy3, Cy5

1. Introduction

Microarrays have rapidly become a widespread tool useful for surveying the levels of mRNA present in cells or tissues at the time of harvest for potentially thousands of genes ***(1)***. DNA microarrays are basically a large number of indi-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

vidual DNA sequences attached to miniature solid substrates, such as glass microscope slides, and are sometimes referred to as "DNA chips." Microarray analysis is conducted by hybridizing DNA chips with one or more fluorescently labeled probes generated from the RNA of the desired tissue(s). Typical microarray analysis employs two RNA-derived probes with different fluorescent dyes (e.g., Cyanine-3 [Cy™3] and Cyanine-5 [Cy5] ***[2]***) hybridized to the attached target DNA elements. Such analysis conveys information of the gene expression patterns between the two tissues or treatments from which the RNA probes were derived. This survey allows for the identification of individual genes that produce contrasting signal intensities across the tissues or treatments as candidates representing differential expression between these tissues or treatments.

Numerous studies have been conducted using microarray analysis and have been comprehensively reviewed (e.g., ***3,4***). In plant systems, microarray technology has been used for a variety of studies, including developmental controls ***(5–8)***, biotic and abiotic stress response ***(9–13)***, nutrient response ***(14)***, and gene family surveys ***(15,16)***. In addition to measuring mRNA levels of sizable gene collections, microarrays have been adapted for genotyping ***(17,18)***, screening for transposon insertions ***(19)***, and protein–protein or protein–ligand interactions ***(20)***. A few microarraying facilities serving the plant-related academic communities have been formed and are accessible via Web sites such as the ZmDB, a maize genomic database (http://www.zmdb.iastate.edu/), the Arabidopsis Microarray Services facility ***(21)***, or the Arabidopsis Functional Genomics Consortium (http://afgc.stanford.edu/). More general microarray resources are also available through many sites such as the Stanford Microarray Database (http://genome-www5.stanford.edu/MicroArray/SMD/index.shtml) and Virginia Tech and North Carolina State University Microarray Technology Resource: Grid It (http://www.bsi.vt.edu/ralscher/gridit/), and with numerous helpful links within. Also, a complete guide for microarraying is available courtesy of Pat Brown's laboratory (http://cmgm.stanford.edu/pbrown/mguide/), including protocols and helpful hints.

Of the two major types of DNA microarrays that have been developed, namely oligonucleotide-based and cDNA microarrays ***(3,22)***, we will discuss the production and use of the cDNA microarray hybridized with two probes (**Fig. 1**). We will describe the preparation of the cDNA products for printing to the solid support, spotting of the DNA chips, producing fluorescently labeled probes using Cy3 and Cy5, pretreatments, hybridizing, and washing of the chips using fairly stringent conditions based in part on the methods of Hedge et al. ***(23)***. Lastly, we will describe the fluorescent scanning, image acquisition, and preliminary data manipulations.

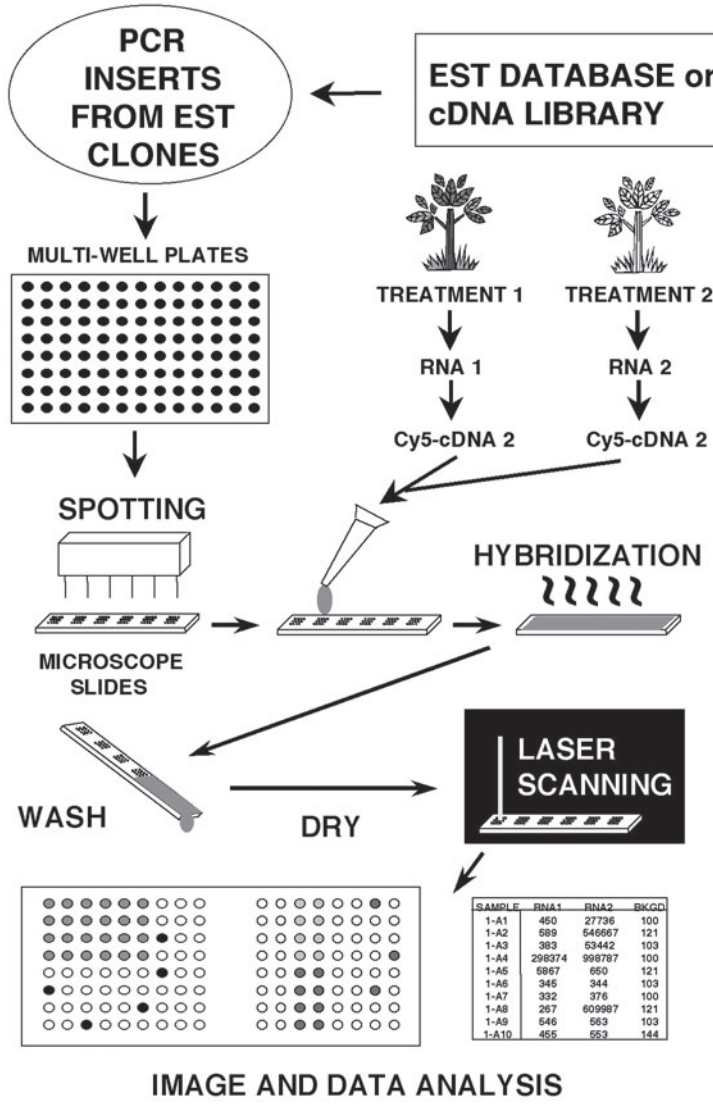


Fig. 1. Schematic of the steps involved in cDNA microarray analysis.

To conduct microarray analysis, access to a robotic printing or spotting system and fluorescent scanners, which can accommodate microscope slides, will be necessary. A DNA printing or spotting system consists of a multi-axis robotic arm with adapted printing heads for various numbers of "quill or solid pens" that dip into the DNA solution present in a multiwell plate and spotting the nanoliter amounts of the DNA samples onto chemically coated glass slides

in an orderly array. An example of less commonly used DNA delivery systems for miniaturized solid supports is ink-jet printers ***(24)***. DNA spotting systems are commercially available from a number of sources, which can be found at links at various Web sites (e.g., http://sgio2.biotec.psu.edu/links/robots.html).

The methods described here will involve the OmniGrid™ Spotter. This system uses a 16-pen head (but can accommodate higher numbers of pens with specialized pen holders) with four 384-well plates, simultaneously, and can print over 23,000 spots of 100–200 μm average diameters on up to 100 microscope slides in less than a 12-h period. The spotter is controlled by a Microsoft® Windows™-based personal computer (PC) running the OmniGrid software and has the flexibility to adjust array designs, spot replications, dipping retention that affects the deposited volumes, and sample tracking–deconvolution capability.

Likewise, several microarray-specific fluorescent scanners are commercially available with information found at numerous Web sites (e.g., http://sgio2.biotec.psu.edu/links/scanners.html). We briefly describe the use of the ScanArray® 5000 with a 20-slide autoloader capability controlled by a Pentium® II NT-based PC with an Ethernet card. Key scanner features are the ability to scan microscope slides with adequate resolution (we typically use 10 μm/pixel), choice of excitation and emission wavelengths useful for flexibility in dye use, and adjustable photomultiplier gain and laser output to improve the quality of acquired slide image data.

2. Materials

2.1. cDNA PCR Fragment Preparations

1. Hotstar *Taq* DNA polymerase kit (Qiagen), including Hotstar *Taq* DNA polymerase, 5× Q-solution, and 10× polymerase chain reaction (PCR) buffer. Store at –20°C.
2. 100 mM dNTP (Roche Molecular Biochemicals) stock solution. Store at –20°C.
3. 100 μM M13 Forward and Reverse primers (Sigma) (the plasmid inserts that we have generated are flanked by the M13F and M13R). Store at –20°C.
4. Distilled water (dH_2O) filtered with 0.1-μm-filtered membrane (Invitrogen).
5. 96-Well PCR plates (VWR Scientific).
6. Nunc™ V-bottom 96-well plate (VWR Scientific).
7. 96-Well multiscreen filter plates (Millipore).
8. Vacuum manifold filtration system (Millipore), controlled pump, or house vacuum system.

2.2. Array Printing

1. Microtiter plate shaker (VWR Scientific).
2. OmniGrid Robotic Spotter (GeneMachines) operated by a Pentium PC running Omnigrid Spotting program.
3. Microarray printing pens (Major Precision Engineering).
4. 50% Dimethyl sulfoxide (DMSO) (Sigma).
5. 384-Well plates (E & K Scientific).
6. Microarray slides (CMT-GAPS-coated slides; Corning). Store slides in a light-tight box in a desiccator.
7. Stratalinker® (Stratagene).
8. Microarray slides box (VWR Scientific).

2.3. Fluorescent Probe Preparation

1. RNaseZAP (Ambion).
2. TRIzol® (Invitrogen). Store at 4°C.
3. Isopropyl alcohol (VWR Scientific).
4. 70% Ethanol.
5. Diethylpyrocarbonate (DEPC)-treated water.
6. 50-mL Oakridge tube (Nalg Nunc International).
7. FastTrack® 2.0 kit for mRNA isolation (Invitrogen), including stock buffer (200 m*M* NaCl, 200 m*M* Tris, pH 7.5, 1.5 m*M* $MgCl_2$, and 2% sodium dodecyl sulfate [SDS], protein/RNase degrader (mixture of proteases), binding buffer (500 m*M* NaCl, 10 m*M* Tris-HCl, pH 7.5, in DEPC-treated water), low salt wash buffer (250 m*M* NaCl, 10 m*M* Tris-HCl, pH 7.5, in DEPC-treated water), elution buffer (10 m*M* Tris-HCl, pH 7.5, in DEPC-treated water), 2 *M* sodium acetate (2 *M* sodium acetate, pH 5.2, in DEPC-treated water), 5 *M* NaCl (5 *M* NaCl in DEPC-treated water), and lyophilized oligo(dT) cellulose.
8. Oligo(dT) (18 to 20-mer; Invitrogen).
9. 1 m*M* Cy3-dCTP and 1 m*M* Cy5-dCTP (Amersham Pharmacia Biotech). Store at –20°C in a light-tight box.
10. 200 U/μL SuperScript II™ kit (Invitrogen) including 5× buffer, 0.1 *M* dithiothreitol (DTT). Store at –20°C.
11. A solution mixture of 2 m*M* dATP, 2 m*M* dGTP, 2 m*M* dTTP, and 1 m*M* dCTP from 100 m*M* dNTP stock solution (Roche Molecular Biochemicals). Store at –20°C.
12. 40 U/μL RNaseOUT (Invitrogen). Store at –20°C.
13. Thermal cyclers (Applied Biosystems or MJ Research).
14. 500 m*M* NaOH.
15. 200 m*M* Free acid morpholinepropanesulfonic acid (MOPS) (Sigma).
16. PCR purification kit (Qiagen) including buffer PB containing chaotropic salt (handle with care), buffer PE, and buffer EB (10 m*M* Tris-HCl, pH 8.5).
17. SpeedVac® (Savant Instruments) or vacuum drying centrifuge with rotor fitted for 96-well plates or 1.5-mL microtubes.

2.4. Hybridization and Wash

1. 5× Sodium chloride sodium citrate (SSC), 0.2% SDS.
2. 4× Hybridization buffer: 20× standard saline citrate (SSC), 0.8% SDS.
3. 20 μg/μL COT1-DNA (Invitrogen).
4. 100% Deionized formamide (Sigma) (toxic, handle with care). Store at –20°C.
5. Oligo(A) 80-mer (Operon Technologies).
6. Slide mailers (Polysciences).
7. Cover slips (Amersham Pharmacia Biotech).
8. Glass jar, cover, tray, and handle (VWR Scientific).
9. MicroDuster III refill and valve accessory (VWR Scientific).
10. 1.0× SSC, 0.2% SDS.
11. 0.1× SSC, 0.2% SDS.
12. 0.1× SSC.
13. Distilled water (dH_2O).

2.5. Scanning and Data Acquisition

ScanArray 5000 scanner, with Pentium II PC running data acquisition and analysis software, ScanArray and QuantArray®, respectively (Packard Biochip Technologies).

3. Methods

3.1. cDNA PCR Fragment Preparations

3.1.1. cDNA Clone Inserts

Plasmid DNA or clones in culture are amplified in 100-μL reactions in 96-well plates using a thermalcycler, according to the following conditions:

1. A reaction master mixture for each 96-well plate includes 6220.8 μL of dH_2O, 960 μL of 10× PCR buffer, 76.8 μL of dNTP mixture (25 m*M* for each), 1920 μL of 5× Q-solution, 38.4 μL of M13 forward primer (100 μ*M*), 38.4 μL of M13 reverse primer (100 μ*M*), and 57.60 μL of HotStar *Taq* DNA polymerase (200 μ/μL).
2. For each clone, add 97 μL master mixture to 3 μL plasmid DNA at 4°C on a thermal cycler and mix well by pipeting up and down 5×.
3. PCR protocol (conditions programmed in an MJ Research thermal cycler): the reaction conditions were 94°C for 15 min, 35 cycles at 94°C for 45 s, 58°C for 45 s, and 72°C for 3 min, and final extension at 72°C for 10 min.

3.1.2. PCR Product Purification Using 96-Well Multiscreen Filter Plates

1. Manually pipet 100 μL PCR products to the Millipore filter plates.
2. Place the filter plate on a vacuum manifold filtration system.
3. Filter at a pressure of 15 in (380 mm) Hg for 15 min.

4. Add 50 µL dH_2O and filter at 15 in (380 mm) Hg for 10 min.
5. Repeat **step 4** 2 twice.
6. Remove the plate from manifold filtration system, place it on a centrifuge, and spin 3 min to remove excess dH_2O.
7. Add 100 µL dH_2O.
8. Place the plate on a shaker and shake vigorously for 15 min to resuspend the DNA.
9. Manually pipet the purified PCR product into a 96-well plate.
10. Dry down the product using the SpeedVac.
11. Seal the plate using a cap mat (VWR Scientific) and store in –20°C for future array preparation.

3.2. Array Printing

1. Add 20 µL of 50% DMSO to each well of the 96-well plate containing dried PCR product (*see* **Note 1**) and mix well by shaking on a microtiter plate shaker for 15 min.
2. Transfer DNA suspension in DMSO to a 384-well plate from the 96-well plate.
3. Centrifuge the plate for 2 min before spotting.
4. Position the plates on the microarray spotter.
5. Label slides (see **Note 2**) with a diamond-tipped pen and remove dust with compressed air.
6. Position the slides (*see* **Note 2**) on the microarray spotter and initiate spotting at 23°C and 45% relative humidity. The OmniGrid spotter allows for up to 100 slides.
7. After spotting, the slides are allowed to air-dry 30 min before UV-crosslink.
8. Cover the 384-well plates containing the DNA/DMSO mixture and store at –20°C.
9. UV-crosslink the slides at 100 mJ using Stratalinker (*see* **Note 3**).
10. Store the slides in a light-tight box in a bench-top desiccator at room temperature (*see* **Note 4**).

3.3. Fluorescent Probe Preparation

3.3.1. Total RNA Extraction Using TRIzol

1. Wash a mortar, pestle, spatulas, and Oakridge tubes with RNaseZAP and rinse 3× with DEPC-treated water before starting extraction (*see* **Note 5**).
2. Add frozen sample tissues to mortar cooled with liquid nitrogen and grind the tissues to a fine powder.
3. Scrape approx 2 g of powder into a 50-mL Oakridge tube containing 20 mL of TRIzol reagent with a spatula precooled with liquid nitrogen and mix well by shaking vigorously for 15 s. No clumps of tissues should be present in the solution.
4. Incubate the samples 5 min at room temperature.
5. Add 4 mL of chloroform to the Oakridge tube containing TRIzol and the tissue sample.
6. Cap the tubes securely and shake them vigorously for 1 min and then incubate at

room temperature for 8 min or until there is a clear separation between the aqueous phase and the organic (red) phase.
7. Centrifuge the tubes at 12,000*g* for 15 min at 4°C.
8. Transfer 13 mL aqueous layer to a new 50-mL Oakridge tube containing 13 mL of 100% isopropanol and mix well.
9. Incubate the tubes at room temperature for 15 min.
10. Centrifuge the tubes at 12,000*g* for 15 min to pellet the total RNA.
11. Discard the supernatant and add 20 mL of 75% ethanol to each tube to wash the RNA pellet. The RNA can be store indefinitely in 75% ethanol at –20°C.
12. Before using RNA samples, centrifuge the tubes at 12,000*g* for 15 min at 4°C and carefully discard as much of the supernatant as possible.
13. Resuspend the pellet in appropriate vol of DEPC-treated water and check RNA quality and quantity before labeling (*see* **Note 6**).

3.3.2. mRNA Isolation Using FastTrack 2.0 Kit from Total RNA

1. Check the Stock Buffer from the FastTrack 2.0 kit. If the stock buffer contains a white precipitate (SDS), heat it to 65°C until fully dissolved.
2. Prepare lysis buffer immediately before use by adding 200 µL of RNase/protein degrader to 10 mL of stock buffer for each intended isolation. Use immediately.
3. Add 1 mg of total RNA in 200 µL of water to freshly prepared 10 mL lysis buffer in a sterile 50-mL centrifuge tube.
4. Heat to 65°C for 5 min, then place immediately on ice for exactly 1 min.
5. Place tube at room temperature and add 650 µL of 5 *M* NaCl, then mix by gentle inversion several times.
6. Add 75 mg oligo(dT) cellulose to the RNA solution.
7. Seal the tube and allow the oligo(dT) to swell for 2 min.
8. Rock the tube gently at room temperature for 15–60 min.
9. Pellet the oligo(dT) at 3000*g* centrifuge for 5 min at room temperature.
10. Remove the supernatant carefully from the resin bed.
11. Resuspend oligo(dT) in 20 mL binding buffer.
12. Centrifuge at 3000*g* for 5 min at room temperature. Remove the supernatant carefully from the resin bed.
13. Resuspend oligo(dT) in 10 mL binding buffer.
14. Centrifuge at 3000*g* for 5 min at room temperature. Remove the supernatant carefully from the resin bed.
15. Resuspend oligo(dT) in 10 mL low salt wash buffer.
16. Centrifuge at 3000*g* for 5 min at room temperature. Remove the supernatant carefully from the resin bed.
17. Repeat **steps 15** and **16** until the buffer is no longer cloudy (4×).
18. Resuspend the oligo(dT) in 800 µL low salt wash buffer.
19. Transfer the oligo(dT) to a spin column.
20. Centrifuge at 5000*g* for 10 s at room temperature.
21. Decant the liquid inside the microcentrifuge tube.

22. Repeat **steps 19–21** to transfer all of the cellulose to the spin column.
23. Wash the oligo(dT) with 550 µL low salt wash buffer.
24. Centrifuge at 5000*g* for 10 s at room temperature.
25. Repeat **steps 23** and **24** until the OD at 260 nm of the "flow-through" is <0.05 (5×).
26. Place the spin column into a new microcentrifuge tube.
27. Resuspend the oligo(dT) in 350 µL elution buffer.
28. Centrifuge at 5000*g* for 30 s at room temperature. Do not decant.
29. Resuspend the oligo(dT) in a second 350 µL elution buffer.
30. Centrifuge at 5000*g* for 30 s at room temperature.
31. Add 105 µL of 2 *M* sodium acetate and 700 µL isopropanol.
32. Freeze on dry ice until solid.
33. Thaw and centrifuge at 12,000*g* for 30 min at 4°C.
34. Remove the supernatant.
35. Wash the pellet with 85% ethanol.
36. Centrifuge at 12,000*g* for 10 min at 4°C, then discard the supernatant.
37. Resuspend the mRNA pellet in 20–50 µL elution buffer.
38. Determine the concentration of the mRNA spectrophotometrically (*see* **Note 6**).
39. Store mRNA at –80°C or use immediately.

3.3.3. Probe Labeling

1. In a 0.2 mL RNase-free thin-wall PCR tube, add 25 µg total RNA (or 250–500 ng mRNA) and 3 µg oligo(dT). Bring to a total vol of 11 µL with DEPC-treated water.
2. Incubate the tubes at 70°C for 10 min and chill on ice for 1 min.
3. Spin the tubes briefly at 3000*g* and place them on ice.
4. Prepare a reaction master mixture for one reaction including 6 µL of 5× SuperScript II buffer, 3 µL of 0.1 *M* DTT, 2 µL of a mixture of 2 m*M* dATP, 2 m*M* dGTP, 2 m*M* dTTP, and 1 m*M* dCTP, 4 µL of 1 m*M* Cy3-dCTP or 1 m*M* Cy5-dCTP (*see* **Note 7**), 1 µL of RNaseOUT (40 µ/µL), and 3 µL SuperScript II (200 µ/µL).
5. Add 19 µL of the master mixture to each reaction tube.
6. Incubate at 42°C for 2.5 h.
7. Add 2 µL of 500 m*M* NaOH and heat at 70°C for 20 min to degrade the RNA.
8. Spin the tubes briefly at 3000*g*. Add 20 µL of 200 m*M* MOPS to neutralize the reaction.

3.3.4. Probe Purification Using PCR Purification Kit

1. Add 5 vol of buffer PB to the probe reaction tube and mixture.
2. Apply the probe to column and spin at 12,000*g* for 1 min.
3. Wash column 3× with 700 µL of buffer PE.
4. Spin column for 1 min after last wash to completely dry column.
5. Elute in 40 µL of 0.1× buffer EB for twice.

3.3.5. Probe Quantification Using UV Spectrophotometer

1. Use 50 µL of the undiluted probe to measure OD at 550 nm for Cy3 and 650 nm for Cy5-labeled probe.
2. For single-stranded probes, use 37 ng/µL for equivalent to 1 OD U at 260 nm.
3. Calculate the total dye incorporation using the following:
   a. Extinction coefficient Cy3 (Ex_{550}) = 150,000 M.
   b. Extinction coefficient Cy5 (Ex_{650}) = 250,000 M.
   c. Total pmol of Cy3 = (OD_{550} × total probe vol)/(Ex550 × 10^{-6}).
   d. Total pmol of Cy5 = (OD_{650} × total probe vol)/(Ex650 × 10^{-6}).
   e. A successful probe preparation should have ≥30 pmol of incorporated Cy dye.

3.4. Hybridization and Wash

3.4.1. Preparation of Slides

1. Place slides in a glass jar containing 5× SSC, 0.2% SDS (preheated to 42°C) and incubate at 42°C for 1 h.
2. Dip the slides 3× in dH_2O at room temperature.
3. Dip the slides 3× in isopropanol at room temperature.
4. Dry slides immediately by blowing compressed air for 15 s.

3.4.2. Hybridization

When performing two-color hybridization, mix an appropriate vol (containing 30 pmol) of each dye according to total dye incorporation calculation described in **Subheading 3.3.5.**, **step 3** and dry down the probe mixture using a SpeedVac.

1. Resuspend the probe in 5 µL dH_2O.
2. Add 7.5 µL of 4× hybridization buffer (20× SSC, 0.8% SDS).
3. Add l µL of 20 µg/µL COT1-DNA.
4. Add 1.5 µL of oligo(A) 80-mer (1 mg/mL).
5. Heat the probe at 92°C for 3 min.
6. Centrifuge at 12,000*g* for 2 min.
7. Add 15 µL of 50% formamide and mix well.
8. Centrifuge at 12,000*g* for 2 min.
9. Add the probe to the slide and cover with 24 × 60 mm cover slips.
10. Place the slide in a slide mailer containing 10 µL of water and wrap the closed lid with parafilm.
11. Incubate the slide mailer at 42°C for 18 h (*see* **Note 8**).

3.4.3. Wash the Slides

Preheat the 1× SSC, 0.2% SDS and 0.1× SSC, 0.2% SDS wash buffers to 55°C. Remove the cover slips under the heated 1× SSC, 0.2% SDS wash buffer. Place the slides in a glass jar (*see* **Note 9**) containing 400 mL of the preheated

solution, shake by a rotary shaker at room temperature for designated time, and then discard the buffer.

1. 1× SSC, 0.2% SDS for 10 min.
2. 0.1× SSC, 0.2% SDS for 10 min.
3. 0.1× SSC, 0.2% SDS for 10 min.
4. 0.1× SSC for 1 min at room temperature.
5. 0.1× SSC for 1 min at room temperature.
6. dH_2O 10 s at room temperature.
7. Dry slides by blowing compressed air (*see* **Note 10**).

3.5. Scanning and Data Acquisition

1. Turn on the scanner and prewarm the appropriate lasers for 15 min.
2. Prescan slides for both of Cy3 and Cy5 and determine the appropriate laser power and photomultiplier tube (PMT) gain (*see* **Note 11**).
3. Scan the slides using predetermined settings.
4. After the scanning is completed, save the image pairs (Cy3 and Cy5 channels) as a 16-bit tagged image format file (TIFF).
5. Use QuantArray software to process image pairs.

3.6. Processing the Image Using QuantArray Software

1. The 16-bit TIFF files can be stored on compact discs at reasonable cost, as they can average over 28 MB in size.
2. Using the QuantArray software, select an appropriate image protocol or create a new protocol using the protocol editor for attributes such as number of subarrays, rows and column distances, spot sizes, background subtraction methods, etc. Select the two TIFF file images for the channels representing the two respective Cy dyes.
3. Move the two TIFF file images into register by the Alignment step.
4. Specify the position of the upper left-most spot of the microarray.
5. Overlay the grid and manually edit the grid row, column, and/or array positions depending on any spotting-mediated distortions.
6. Determine the center of each spot and surrounding area by using the Locate Spot step. We typically check the nominal locations box if we were careful in aligning the grid in **step 5** above to more rapidly create the grid.
7. Flag any spot using the Ignore Filter in both channels that show spot distortions, smearing, or general noncircular signals (**Fig. 2**). For more information on microarray distortions, *see* ref. ***(25)***.
8. Select acquire data and save as a spreadsheet. Upload data in any appropriate database. The user may select one of several background subtraction methods (*see* **Note 12**). The data output can be analyzed by QuantArray using one of several tools (*see* **Note 13**).
9. QuantArray provides Excel® macros that assist in normalizations, so that data from each channel for a single DNA chip is adjusted for comparisons with the

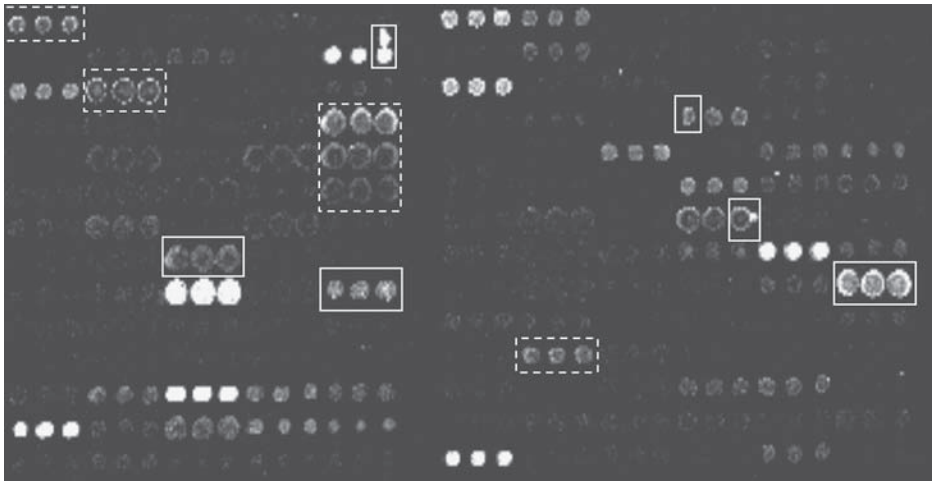


Fig. 2. Examples of spots of marginal quality that were either rejected or accepted. We typically reject spots that show gross amorphous shapes, "spill over" into neighboring spots, or other obvious flaws as shown by a few examples in small white boxes (also *see* ref. ***25***). We also avoid spots that contribute to correlation coefficient values >0.15% in replicates as depicted by the triplet spots in solid white boxes. We have noted that some spots generated by the microarray system we described (see text) were produced as rings rather than solid spots. These were generally accepted if they generated coefficient of variations (CVs) of <0.15%. Examples of triplicate spots with ring-shape pattern designated by the dotted white boxes produced a CV of 10% or less, as do the remaining replicates shown.

other channel or other DNA chips. The three normalization macros are as follows (*see* **Note 14**):

a. Based on the mean of the total data for a specific channel.
b. Based on the median value of the total data for a specific channel.
c. Based on values for a selected set of cDNA spots (e.g., corresponding to control or housekeeping genes).

10. In designing a microarray experiment, both biological and technical replications are necessary to generate high quality data (*see* **Note 15**).
11. To confirm the differential expression of specific candidate genes based on the microarray data, we typically conduct conventional methods such as RNA gel blot analysis, reverse transcription polymerase chain reaction (RT-PCR) techniques, etc., using gene-specific probes (*see* **Note 16**).
12. To obtain the highest quality and reproducible microarray data, there are three very critical steps that need extra attention. These include avoiding poor quality slides, avoiding poor handling during preparations for hybridizations (*see* **Note 17**), and avoiding dust contamination (*see* **Note 18**).

4. Notes

1. The concentration of the target DNA (cDNA PCR fragments) should be approx 200 fmol/μL. If the concentration of the target is lower, program the spotter to dip twice during a complete spotting run.
2. Always wear gloves when handling the slides.
3. Do not exceed 100 mJ for the UV crosslinking step to avoid reduction in signal strength of final data.
4. Always keep the slides in a light-tight box in a desiccator when not in use, because DMSO is light sensitive.
5. When making RNA, precautions should be taken to ensure that all RNase activity is destroyed or avoided. Gloves should always be worn and DEPC-treated solution should be used wherever possible.
6. Determine the RNA concentration and purity by spectrophotometric measurement and running an aliquot (5 μg for total RNA and 300 ng for mRNA) on a formaldehyde gel consisting of 1.2% SeaKem® agarose, 1× MOPS, and 4% formaldehyde running with 1× MOPS buffer. Stain gel with ethidium bromide (2 μg for each sample). Determine the RNA concentration by using the following formula: $[RNA] = (A_{260})(0.04\ \mu g/\mu L)(D)$. D is the dilution factor.
7. Avoid exposing the Cyanine dyes to the light as much as possible.
8. Limit the hybridization time to <20 h to avoid high backgrounds.
9. For the posthybridization washes, wash the slides in a very clean container (a glass jar is recommended) and cover it with aluminum foil to avoid light.
10. Dry slides with compressed air immediately. Do not let slides to air-dry or spotting will occur.
11. When scanning the slides, select the appropriate laser power and PMT gain. The range for laser power is 0–100% and PMT gain is 33–100%. Increasing laser power and PMT gain improves sensitivity, but also increases background noise and leads to signal saturation. We typically choose 80% laser power and 80% PMT gain for most scans using the ScanArray 5000.
12. We observed that using local background subtraction methods sometimes leads to erroneous and nonreproducible data. We instead rely on a set of blank spots for correcting the raw values. We use the mean plus one standard deviation value of at least 200 blank spots per DNA chip (i.e., no DNA solution present in randomly chosen wells, designated as blanks, of the microtiter plates used in printing) to ascertain average backgrounds levels.
13. There are many useful tools in analyzing the data output, such as observing the morphology of individual spots, signal intensities of spots with accompanied statistics, and several graphic modes for comparing the two channels of individual and collective sets of spots.
14. Following the initial background-subtracted data collection from the TIFF image files, numerous methods of normalizing, transforming, and analyzing the data have been reported (*see* refs. within ***[1]***). Of the three methods of normalization mentioned above, we typically use the "all data" method. This is based on observations from a large number of treatment comparisons of signal intensities cor-

responding to the majority of transcripts from a variety of cells or tissues that do not show significant differences between treatments (data not shown).

15. Replication is important for producing useful data. Since significant variations in the initial RNA preparation, probes, and hybridization reactions are likely, we usually conduct a minimum of three replicated slides as recommended by Lee et al. ***(26)***. Depending on experimental design, we typically employ dye-swap hybridizations, where we hybridize the first microarray chip with Cy3-labeled probe-1 and Cy5 probe-2 while the duplicate chip is hybridized with dye-swapped probes (i.e., Cy3 probe-2 and Cy5 probe-1). For each cDNA entry on the chip, we use the sum of the two dye-swap chips for ensuing data analysis.
16. Since cDNA microarrays are dependent on hybridizing reversed-transcribed probes to immobilized cDNA targets, the potential for cross hybridization is possible and adversely affects the assessment of gene-specific expression levels. A few studies have been conducted addressing this issue ***(15)*** and have concluded that homology of less than 80% sequence identity between cDNA entries on a chip causes very little cross hybridization. The methods outlined above are relatively stringent conditions and reduces or eliminates cross hybridization with sequences <75% identical that are >500 bp long (data not shown).
17. To avoid poor quality slides, we prescan each slide with the Cy3 channel and discard those that show any distortions or patterns, which will affect the spotting and cDNA retention during the hybridizations. The second major source of problems arises during the placement of the cover slip onto the slide after adding the hybridization solution. While placing the cover slip, it is easy for bubbles to form or to smear the cDNA spots. To remove bubbles, small amount of pressure can be placed on the cover slip after it is fully in place to carefully force the bubble to the edge and out from under the cover slip. Smearing can be avoided by preventing the subsequent sliding of the cover slip once it is in place.
18. A third major source of problems is contaminating dust and minute debris on the microarrays that will obscure or otherwise compromise the data. We conduct all operations involving the slides in a HEPA-filtered room under positive air-flow pressure. We also avoid having corrugated or storage boxes and excess paper goods in the same room and endeavor to keep the room very clean. We also use compressed air cans to help blow any remaining dust off of slides before spotting and scanning, the spotter itself, and any working surfaces and vessels used in washing the slides.

References

1. Lockhart, D. J. and Winzeler, E. A. (2000) Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836.
2. Yu, H., Chao, J., Patek, D., Mujumdar, R., Mujumdar, S., and Waggoner, A. S. (1994) Cyanine dye dUTP analogs for enzymatic labeling of DNA probes. *Nucleic Acids Res.* **22,** 3226–3232.
3. Lockhart, D. J., Dong, H., Byrne, M. C., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680.

4. Lipshutz, R. J., Morris, D., Chee, M., et al. (1995) Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques* **19,** 442–447.
5. Chatterjee, A. and Roux, S. J. (2000) *Ceratopteris richardii*: a productive model for revealing secrets of signaling and development. *J. Plant Growth Regul.* **19,** 284–289.
6. Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiol.* **124,** 1570–1581.
7. Hertzberg, M., Aspeborg, H., Schrader, J., et al. (2001) A transcriptional roadmap to wood formation. *Proc. Natl. Acad. Sci. USA* **98,** 14732–14737.
8. Whetten, R., Sun, Y. H., Zhang, Y., and Sederoff, R. (2001) Functional genomics and cell wall biosynthesis in loblolly pine. *Plant Mol. Biol.* **47,** 275–291.
9. Desikan, R., A.-H.-Mackerness, S., Hancock, J. T., and Neill, S. J. (2001) Regulation of the *Arabidopsis* transcriptome by oxidative stress. *Plant Physiol.* **127,** 159–172.
10. Kawasaki, S., Borchert, C., Deyholos, M., et al. (2001) Gene expression profiles during the initial phase of salt stress in rice. *Plant Cell* **13,** 889–905.
11. Seki, M., Narusaka, M., Abe, H., et al. (2001) Monitoring the expression pattern of 1300 *Arabidopsis* genes under drought and cold stresses by using a full-length cDNA microarray. *Plant Cell* **13,** 61–72.
12. Maleck, K., Levine, A., Eulgem, T., et al. (2000) The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nat. Genet.* **26,** 403–410.
13. Schenk, P. M., Kazan, K., Wilson, I., et al. (2000) Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. *Proc. Natl. Acad. Sci. USA* **97,** 11655–11660.
14. Wang, R., Guegler, K., LaBrie, S. T., and Crawford, N. M. (2000) Genomic analysis of a nutrient response in *Arabidopsis* reveals diverse expression patterns and novel metabolic and potential regulatory genes induced by nitrate. *Plant Cell* **12,** 1491–1509.
15. Xu, W., Bak, S., Decker, A., Paquette, S. M., Feyereisen, R., and Galbraith, D. W. (2001) Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana. Gene* **272,** 61–74.
16. McGonigle, B., Keeler, S. J., Lau, S. M., Koeppe, M. K., and O'Keefe, D. P. (2000) A genomics approach to the comprehensive analysis of the glutathione *S*-transferase gene family in soybean and maize. *Plant Physiol.* **124,** 1105–1120.
17. Nouzova, M., Neumann, P., Navratilova, A., Galbraith, D. W., and Macas, J. (2001) Microarray-based survey of repetitive genomic sequences in *Vicia* spp. *Plant Mol. Biol.* **45,** 229–244.
18. Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001) Diversity arrays: a solid-state technology for sequence information independent genotyping. *Nucleic Acids Res.* **29,** E25.
19. Mahalingam, R. and Fedoroff, N. (2001) Screening insertion libraries for mutations in many genes simultaneously using DNA microarrays. *Proc. Natl. Acad. Sci. USA* **98,** 7420–7425.

20. Stoll, D. F., Schrenk, M., Traub, P. C., Vohringer, C. F., and Joos, T. O. (2002) Protein microarray technology. *Front. Biosci.* **7,** C13–C32.
21. Wisman, E. and Ohlrogge, J. (2000) *Arabidopsis* microarray service facilities. *Plant Physiol.* **124,** 1468–1471.
22. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470.
23. Hegde, P., Qi, R., Abernathy, K., et al. (2000) A concise guide to cDNA microarray analysis. *BioTechniques* **29,** 548–556.
24. Okamoto, T., Suzuki, T., and Yamamoto, N. (2000) Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat. Biotechnol.* **18,** 438–441.
25. Brown, C. S., Goodwin, P. C., and Sorger, P. K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. USA* **98,** 8944–8949.
26. Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97,** 9834–9839.

23

Open Architecture Expression Profiling of Plant Transcriptomes and Gene Discovery Using GeneCalling® Technology

Oswald R. Crasta and Otto Folkerts

Summary

The recent rapid developments in genomics tools, technologies, and bioinformatics have revolutionized gene expression analysis. It is now routine to measure gene expression modulation at the genomic level. GeneCalling® technology is an open architecture system capable of assaying more than 95% of genes expressed in a tissue. Unlike the closed systems, GeneCalling is not dependent upon an existing sequence or clone database. GeneCalling uses as low as 50 pg of the cDNA from samples and identifies cDNA fragments that are differentially modulated within a set of samples. With the use of 96 pairs of restriction enzymes, more than 30,000 cDNA fragments are routinely assayed to identify those that are differentially modulated. Specific processes, such as SeqCalling™, Trace Poisoning, and GeneCall Poisoning, are set up to not only confirm the known genes, but also to clone and analyze unknown and novel genes that have an interesting expression profile. GeneCalling has been successfully applied to expression profiling of several plant and fungal species, and resulted in identification and characterization of genes that are useful in commercial applications towards improving agriculturally important traits in plants.

Key Words

transcript profiling, transcriptome, differential gene expression, GeneCalling, SeqCalling, Trace Poisoning, GeneCall Poisoning

1. Introduction

During the last decade, there has been tremendous progress in the development of novel technologies and bioinformatics tools used for genome-wide discovery of genes and deciphering their function. The term "transcript profil-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

ing" refers to the analysis of the amount of mRNA messages of specific genes produced in specific tissue samples. Association of genes to specific phenotypic functions using transcript profiling technologies relies on the assumption that transcriptional regulation of genes plays a key role in expression of phenotypes. Rapid growth in the number and type of technologies and the development of powerful bioinformatic tools have transformed the expression analysis from one gene at a time ***(1)*** to simultaneous profiling of thousands of genes in hundreds of tissue samples representing different treatments or conditions associated with the experiment ***(2)***. The ability to profile almost all genes expressed in a tissue led to the use of the term "transcriptome," which refers to the sum total of the expressed genes in a genome in that tissue under specific conditions.

The term "differential gene expression" refers to a context-dependent variation in expression of the transcripts in a number of samples that are of interest in an experiment. In most experimental cases, the differential expression of genes in meaningful sample comparisons forms the focus of the analysis, rather than merely the determination and analysis of the genes expressed in such samples. High-throughput differential expression technologies provide a genome-wide view of the changes in transcriptional regulation and, hence, enhance the ability to associate genes to phenotypes. Mainly, there are two types of differential gene expression technologies. The first type are the closed systems, in which the discovery of genes depends upon the availability and use of the known or sequenced genes, while the second type are the open systems, which identify the genes that are differentially expressed in samples or treatments independent of the availability of known or sequenced genes. Essentially, the outcome of the experiments in both systems is global differential expression analysis of genes by combining information from two entities: *(i)* a list of genes; and *(ii)* their expression relationship in a relatively large number of samples of experimental interest. The main difference between the two systems is deciding which entity is taken as a reference point. In a closed system, a gene list is taken as a reference, and the expression data for these genes is obtained for a set of samples. Whereas in the open system, the set of samples is taken as the reference, while the list of genes is open-ended with an upper limit of all the genes expressed in a tissue. The most desirable outcome in differential gene expression experiments and analysis is to identify all the genes that are differentially expressed between any two treatments in the transcriptome.

In this chapter, we describe the application of GeneCalling®, which is an open system of transcript profiling, for functional analysis of genomes. Expression profiling using GeneCalling technology is independent of the organism, species, experimental design, treatment, and more importantly, independent of the availability and quality of a sequence database ***(3)***. This

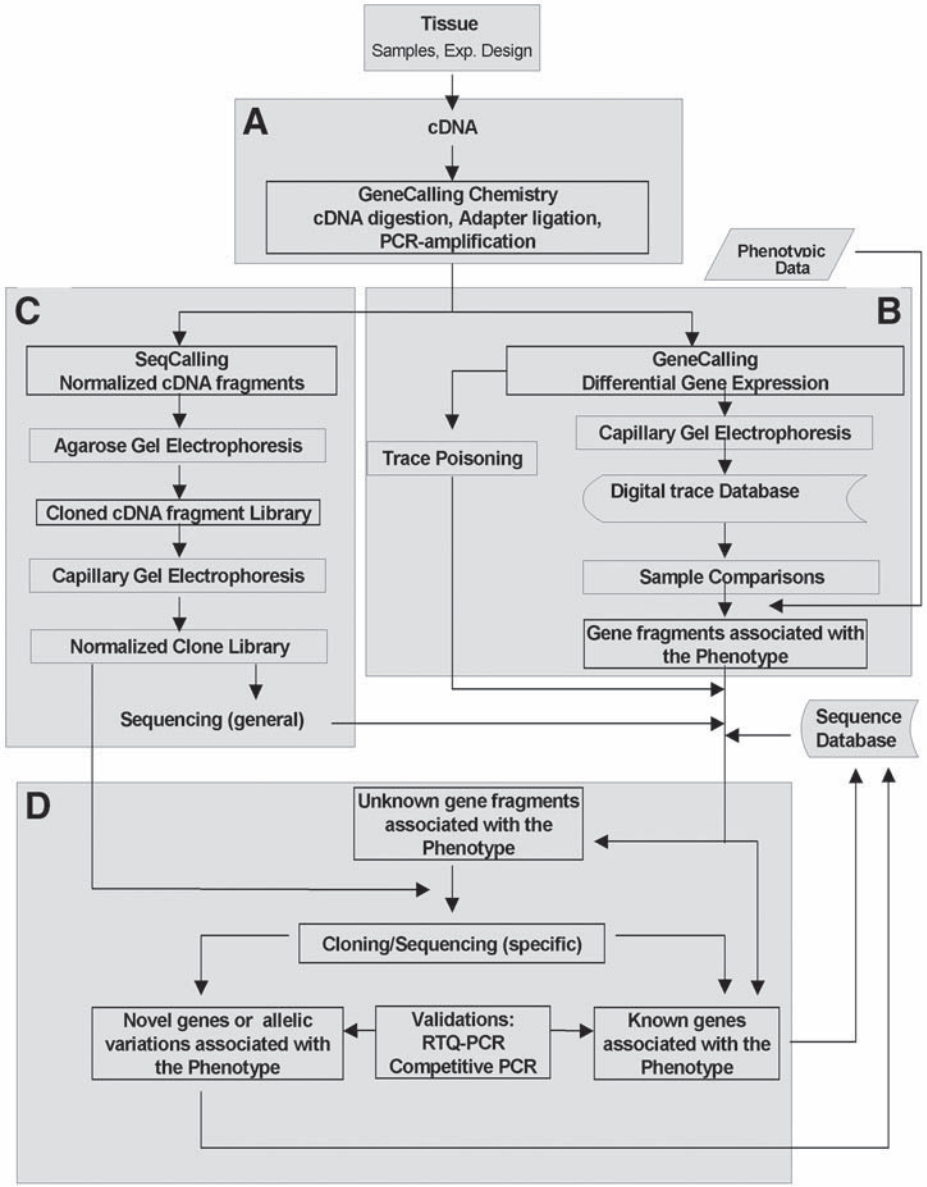


Fig. 1. Overview of the GeneCalling process and related technologies. **(A)** Sample preparation and processing. **(B)** GeneCalling transcript profiling. **(C)** SeqCalling: cloning and sequencing of unknown gene fragments. **(D)** Data integration, analysis, and gene discovery.

technology has been successfully used in several organisms with numerous experimental designs *(2)* targeted for discovery of known and unknown or novel genes that are differentially expressed. In this chapter, we provide a basic and general description of the GeneCalling technology and its application in the discovery of genes and their function at the genomic level. An overview of the GeneCalling process is given in **Fig. 1**. The description of GeneCalling is broken up in the four main steps defined by the workflow of the process.

GeneCalling has been applied for transcriptional profiling in several plants including maize, wheat, soybean, sunflower, canola, and tomato. This technology has also been applied to the discovery of genes and pathways in eukaryotic microbes relevant to crop improvement through plant genetic modifications. A summary of the transcriptional profiling in several organisms is given in **Table 1** *(4–12)*.

2. Materials

2.1. Sample Preparation

1. RNA extraction: Trizol® (Life Technologies) or Tripure reagent (Roche Molecular Biochemicals) and fluorometry with OilGreen® (Molecular Probes).
2. Poly(A)+ RNA: about 100 μg total RNA and oligo(dT) magnetic beads (PerSeptive Biosystems).
3. First strand cDNA synthesis (limit this to materials, number of units should be in **Subheading 3.**): poly(A)+ RNA using oligo$(dT)_{25}$V (V = A, C, or G) (Amitof Biotech) and SuperScript® II reverse transcriptase (Life Technologies).
4. Second strand cDNA synthesis: DNA ligase, DNA polymerase I, RNase H, T4 DNA polymerase (all Life Technologies), and arctic shrimp alkaline phosphatase (USB).
5. cDNA purification and quantification: phenol–chloroform (1:1), ethanol, and fluorometry with PicoGreen® (Molecular Probes).

2.2. Sample Processing

The GeneCalling reactions:

1. 96 Pairs of 6-nucleotide (nt) recognizing restriction enzymes (RE) *(3)* with specific buffers.
2. Two primers–adapters. One of the primers is labeled with the fluorescent FAM label (J primer), while the other is labeled using Biotin (R primer) (Operon Technologies).
3. T4 DNA ligase (Invitrogen).
4. Polymerase chain reaction (PCR) amplification reagents: 10 m*M* dNTP, 10× TB buffer (500 m*M* Tris, 160 m*M* $[NH_4]_2SO_4$, 20 m*M* $MgCl_2$, pH 9.15), Klentaq (Clontech Laboratories), *Pfu* (Stratagene), and water.

Table 1
Examples of the Differences Seen in Different GeneCalling Experiments

Organism	Tissue	No. of treatments	No. of RE pairs	No. of cDNA fragments assayed	Proportion (%) of differentially modulated cDNA fragments		Reference
					Minimum	Maximum	
Maize	BMS cells	9	68	>19,000	0.4	2.0	***(4)***
Maize	Root	4	42	>13,000	1.7	2.4	***(10)***
Maize	Leaf	6	48	>13,500	3.4	19.0	***(12)***
Maize	Embryo	2	89	>32,000	0.3	0.3	***(11)***
Tomato	Leaf	6	96	>34,000	NA	13.0	***(5)***
Black yeast	Mycelia	2	93	>27,000	3.0	3.0	***(9)***

2.3. GeneCalling Transcript Profiling

1. PCR product purification and denaturation reagents: streptavidin beads (CPG), buffer 1 (3 *M* NaCl, 10 m*M* Tris-HCl, 1 m*M* ethylene diaminetetraacetic acid [EDTA], pH 7.5), buffer 2 (10 m*M* Tris-HCL, 1 m*M* EDTA, pH 8.0), buffer 3 (80% [v/v] formamide, 4 m*M* EDTA, 5% TAMRA- or ROX-tagged molecular size standards [Applied Biosystems]), 6 *M* urea.
2. Electrophoresis of purified PCR product: MEGABACE capillary electrophoresis (Molecular Dynamics).
3. Gel interpretation: Open Genome Initiative (OGI) software ***(3)*** (CuraGen Corporation).
4. Data analysis: GeneScape software ***(13)*** (CuraGen Corporation).

3. Methods

3.1. Sample Preparation and Processing

1. Total RNA is extracted from the tissue with Trizol or Tripure isolation reagent according to the manufacturer's instructions.
2. The quality of the total RNA is tested by spectrophotometry and formaldehyde gel electrophoresis, and the total RNA yield is estimated by fluorometry with OilGreen.
3. Poly(A)+ RNA is isolated from about 100 mg total RNA using oligo(dT) magnetic beads.
4. The first-strand cDNA is prepared from 1.0 μg of poly(A)+ RNA with 200 pmol oligo$(dT)_{25}$V (V = A, C, or G) using 400 U of SuperScript II reverse transcriptase.
5. The second-strand synthesis is done at 16°C for 2 h after addition of 10 U of *Escherichia coli* DNA ligase, 40 U of *E. coli* DNA polymerase, and 3.5 U of *E. coli* RNase H. After incubation with T4 DNA polymerase (5 U) for 5 min at 16°C, arctic shrimp alkaline phosphatase (5 U) is added and incubated at 37°C for 30 min.
6. cDNA is purified by phenol–chloroform extraction, and its yield is estimated using fluorometry with PicoGreen. The cDNA quality is tested by subjecting the samples to GeneCalling chemistry (*see* **Note 1**).

3.2. GeneCalling Transcript-Profiling: Modulated Gene-Fragments

The GeneCalling chemistry of cDNA samples is performed in 96-well or 384-well plates. In the 384-well format, 50 pg of cDNA is enough per reaction, while the 96-well format requires 1 ng. The materials for GeneCalling chemistry listed here are for the 96-well format ***(4,13)***.

Each GeneCalling reaction involves:

1. Digestion of 1 ng of cDNA in 50 μL by a unique pair of REs that leave 4 bp 5' cohesive ends. The manufacturer's instructions and buffers are used for the digestion. This reaction is repeated 96 times with separate pairs of REs.

2. Ligation of cDNA fragments with compatible amplification tags at 16°C for 1 h in 10 m*M* ATP, 2.5% polyethylene glycol (PEG), 10 U T4 DNA ligase, and 1× ligase buffer.
3. PCR amplification of the ligated cDNA fragments is done using two primers specific to the ligated tags at each end of the cDNA fragments ***(5)***. One of the primers used for PCR amplification is labeled with the fluorescent FAM label (J primer), while the other is labeled using Biotin (R primer).
4. PCR amplification is performed after addition of 2 μL 10 m*M* dNTP, 5 μL 10× TB buffer (500 m*M* Tris-HCL, 160 m*M* $[NH_4]_2SO_4$, 20 m*M* $MgCl_2$, pH 9.15), 0.25 μL Klentaq:*Pfu* (16:1), 32.75 μL water.
5. PCR amplification is carried out for 20 cycles (30 s at 96°C, 1 min at 57°C, 2 min at 72°C), followed by 10 min at 72°C.
6. PCR product purification is done using streptavidin magnetic beads, washed twice with buffer 1. Buffer 1 (20 μL) was mixed with the PCR product for 10 min at room temperature, separated with a magnet, and washed once with buffer 2, dried, and resuspended in 3 μL of buffer 3.
7. The purified PCR product is denatured using 6 *M* urea and analyzed using MEGABACE capillary electrophoresis.
8. Electrophoresis data is processed using the OGI software.
9. A quality control step is done to check for low signal-to-noise, poor peak resolution, missing ladder peaks, and lane-to-lane bleed. The data that pass this quality control (QC) criteria is submitted as point-by-point length vs amplitude addresses to an Oracle 8 database.
10. Appropriate replications and repetitions are included to evaluate for technological and biological variation (*see* **Note 2**).
11. GeneScape software ***(13)*** is used to normalize the data and to identify gene fragments that are differentially modulated. Differential expression analysis (DEA) is performed either by pairwise comparison of the treatments or simultaneous multisample comparison of all treatments. In both cases, the traces from different treatments are normalized using a scaling algorithm. Following the normalization, the mean and standard deviation (biological variation) of the peak intensities of cDNA fragments for different treatments are used to identify differentially modulated fragments (with a typical N-fold threshold of ±1.5 and *P* values ≤0.01) (*see* **Note 3**).

3.3. GeneCalling Transcript Profiling: Known and Novel Genes

Several attributes of the cDNA fragments measured in GeneCalling are compared and matched to those of the sequences in an available sequence database to associate the cDNA fragments with known genes and to detect unknown and potentially novel gene fragments. These attributes are described below:

1. GeneCalling: to associate the modulated cDNA fragments with known gene sequences, the known RE pair sequence information, combined with the measured length of the fragment are used. The sequence database is used to generate

the list of predicted gene fragments, by *in silico* digestion of the sequence database with the same pairs of REs used in GeneCalling. The candidate genes associated with the modulated gene fragments are obtained by matching the RE pairs between the predicted and modulated gene fragments, such that the difference in the fragment sequence length and the electrophoretic size of the modulated fragment is within 1.5 bp.

2. Trace Poisoning: Trace Poisoning is a modified and general version of the competitive PCR process, GeneCall Poisoning ***(4,5,13,14)***. In Trace Poisoning, the PCR product from the GeneCalling reaction is reamplified under the original conditions with the addition of one of 16 unlabeled oligonucleotide primers. The unlabeled primers are identical to one of the labeled primers (J or R primer) except that the 3' ends are extended to include one (A, T, G, or C) or two (NA, NT, NG, or NC) nucleotides immediately following the unique restriction site at each end (J and R). The traces from such reamplification will be identical to the original traces (no unlabeled primers), except that all the peaks pertaining to a subset of the cDNA fragments, whose nucleotides adjacent to the 6-bp RE recognition site are complementary to the additional nucleotides present in the unlabeled primer, are ablated or reduced in intensity. Thus, the specific nucleotides at the first or second position immediately adjacent to the RE site are identified by observing which of the four primers ablate the peak or reduce the intensity of a given cDNA fragment in a trace. Thus, Trace Poisoning identifies up to four additional nucleotides immediately adjacent to the RE sites at both ends of the cDNA fragment. An example of identification of additional nucleotides in cDNA fragments using Trace Poisoning is given in **Fig. 2**.
3. SeqCalling™: Cloning and sequencing of unknown gene fragments with SeqCalling is complementary to GeneCalling, and provides additional accuracy of detecting known gene sequences. More importantly, SeqCalling supplements GeneCalling by providing a powerful method of detecting and efficiently cloning fragments corresponding to unknown and novel gene sequences that are not represented in the available sequence database. However, SeqCalling is not required if the organism has a fully sequenced genome or if the goal of the experiment is to discover only known genes. SeqCalling involves creating a normalized clone library using the PCR product of the GeneCalling reaction. The PCR product of the GeneCalling reaction (pooled from all treatments) from each of 96 RE pairs is subjected to electrophoresis using MetaPhor® gels (Cambrex Corporation). Based on migration of standard size ladders present in adjacent lanes during the electrophoresis, the experimental lanes in the gel are size-fractionated into 48 sections between 40 and 450 bp. The eluates of these fractions are cloned into standard vectors, and up to 96 clones from each of the fractions are selected and stored. All these clones (96 RE pairs × 48 fractions × up to 96 clones) are then PCR-amplified using the standard vector primers and subjected to the same precise sizing (±0.2 bp) as used for GeneCalling transcript profiling. The clone library is normalized by grouping the clones ranging in size within 0.2 bp. One or two representative clones from each size group are subjected to sequencing to

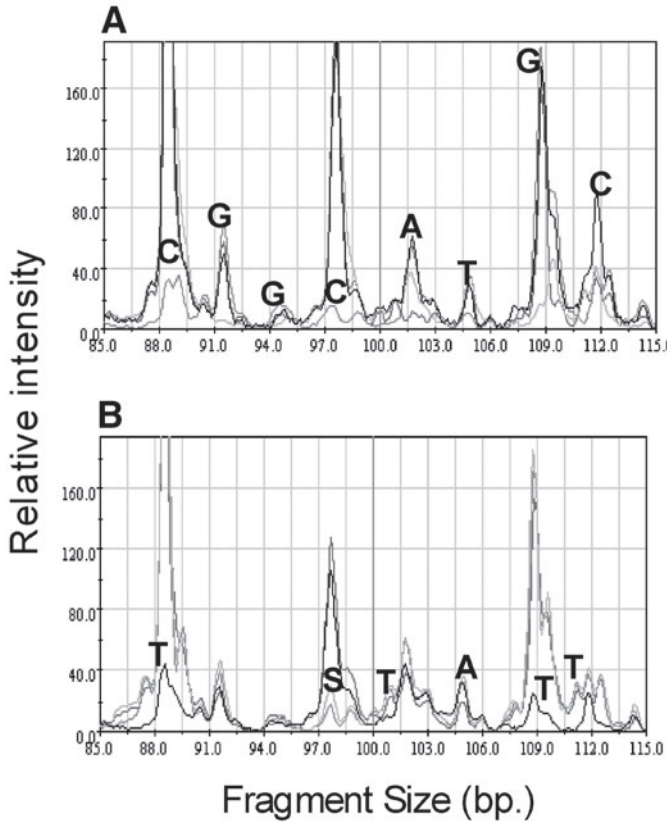


Fig. 2. The Trace Poisoning process yields additional sequence information for bands detected in GeneCalling. **(A)** Identification of the nucleotides in R2 position (second nucleotide from the RE site at the R end). **(B)** Identification of the nucleotides in J2 position (second nucleotide from the RE site at the R end). The identified nucleotides of the corresponding cDNA fragments are indicated next to each peak in the digital images.

generate a comprehensive sequence database covering more than 80% of all the cDNA fragments detected in GeneCalling. This process can be used either to produce an unbiased database prior to, and independent of, the GeneCalling experiments or a biased database of selected modulated fragments after the GeneCalling data has been obtained. In the latter case, only those clones with sizes within 0.2 bp from that of the modulated cDNA fragments in the GeneCalling experiment, which have a desirable expression profile, are subjected to sequencing.

3.4. Data Integration, Analysis, and Gene Discovery

The putative association of gene fragment to known gene is then converted into an overall significance value, T_j, using the following formula:

$T_j = 1 - (F1*F2*F3*\ldots Fi)$

where Fi = false positive rate of matching of i^{th} cDNA fragment with j^{th} gene as described in **Subheading 3.3**.

Trace Poisoning and SeqCalling processes provide additional attributes to associate cDNA fragments to known genes with significance values (*see* **Note 4**).

3.5. Discovery of Unknown and Novel Genes

As mentioned before, GeneCalling is an open architecture system capable of assaying >95% of genes expressed in any given tissue. This provides an opportunity to discover unknown and novel genes that are co-regulated with the phenotype of interest. Often, several cDNA fragments are not associated with known sequences. The proportion of such cDNA fragments depends on the coverage of the transcriptome in the available sequence database, the quality and length of the sequences, and the allelic variations that are present in the processed samples as compared to the available sequences. All of these cDNA fragments are of potential interest in evaluating their association with the phenotype of interest. The creation of a normalized SeqCalling clone library will significantly increase efficiency (both time and cost) of detecting the sequences associated with such unknown or novel cDNA fragments (**Fig. 1**) ***(15)***. A targeted sequencing approach can be taken to use the SeqCalling data efficiently (*see* **Note 5**).

3.6. Confirmation and Validation of Gene Expression Profiles

The association of a differentially modulated gene fragment to a known gene, as identified using the steps described above, is confirmed by subjecting a representative set of the cDNA fragment and matched gene sequence pairs to GeneCall Poisoning, a specific competitive PCR confirmation process ***(4,5,13)***. GeneCall Poisoning is designed to unambiguously confirm the association between a specific cDNA fragment and the gene sequences to which it is associated in the database. This process is identical to Trace Poisoning, except that only one unique unlabeled primer is used in the reamplification process.

1. This unique unlabeled primer is designed to extend into the specific sequence of the candidate gene fragment.
2. The unlabeled primer is between 20–25 nt long, consisting of one of the 6 bp RE sites of the predicted match, plus an additional 14–19 nt of sequence complemen-

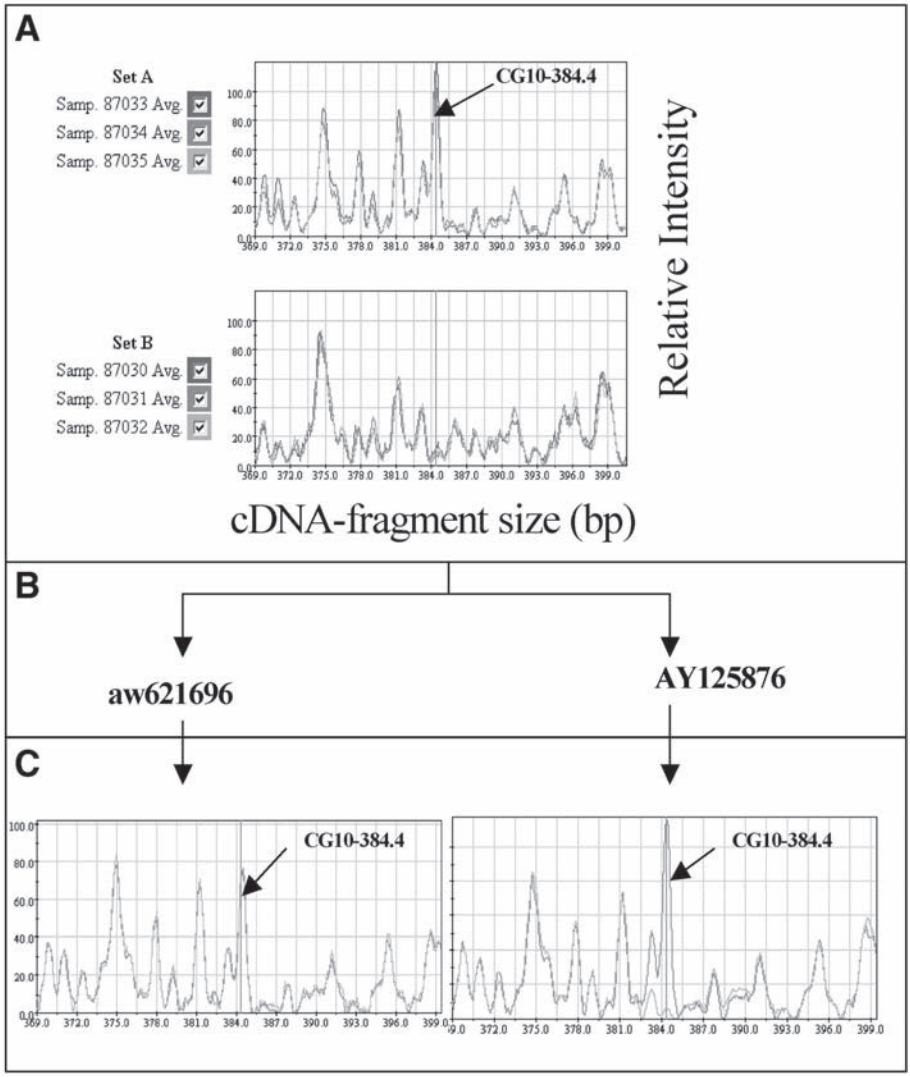


Fig. 3. The GeneCall Poisoning confirmation process unambiguously assigns bands to genes in the sequence database. The figure demonstrates positive (true) and negative (false) associations between cDNA fragments and the known gene sequence. **(A)** A cDNA fragment (CG10–384.4), shown by the arrow, was detected to be induced in the leaf of the PtoR line infected with the pathogen (Set A) as compared to the control uninfected line (Set B) ***(5)***. **(B)** Two examples of sequences (GeneBank® accession numbers) associated with the cDNA fragment by GeneCalling. **(C)** Confirmation of the association by GeneCall Poisoning. The reamplification of the control trace using the unlabeled probe specific to the associated sequences is shown in panel B. The GeneCall Poisoning results confirmed that the fragment CG10-384.4 was part of the gene sequence AY125876 (true positive), but not aw621696 (false positive).

tary to the corresponding region from the associated gene sequence. If the specific cDNA fragment is, in fact, derived from the associated gene sequence, the unlabeled primers will anneal to the cDNA fragments much more efficiently than the FAM-labeled primer, resulting in competition with the labeled fragment in the amplification reaction and ablation of the peak of the cDNA fragment (positive confirmation). If the association between the cDNA fragment and the gene sequence is false, the peak will remain intact during the reamplification (negative confirmation). **Figure 3** shows examples of the GeneCall Poisoning confirmation process with positive (true) and negative (false) associations between cDNA fragments and the known gene sequence.

Although GeneCall Poisoning provides a definitive gene identity to a cDNA fragment, it is not necessary to confirm all differentially cDNA fragments in this manner (*see* **Note 6**).

4. Notes

1. It is critical to get good quality cDNA for accurate profiling of the samples. The cDNA is tested by subjecting each samples to GeneCalling reaction using 4–8 pairs of RE and evaluating the quality and consistency of the Trace profiles.
2. Typically, each experimental treatment is replicated with three independent RNA samples to cover the biological variation, and each RNA sample is repeated three times in GeneCalling chemistry to cover any variation introduced by the technology. In a typical experiment, each RNA sample is processed with up to 96 separate RE pairs, for an estimated coverage of more than 95% of the transcriptome ***(13)***.
3. The number or the percentage of genes that are differentially modulated in a typical experiment depends upon many factors, such as treatment, organism, tissue, and developmental stage. The proportion of differentially modulated cDNA fragments to the total detected varies from 0–15% in a typical experiment. **Table 1** shows examples of the differences seen in different GeneCalling experiments.
4. The effect of Trace Poisoning on associating individual cDNA fragments with known gene sequences is shown in **Fig. 4**. GeneCalling Poisoning data from several GeneCalling experiments in different organisms were analyzed (Crasta, unpublished data). Data from a total of 1564 GeneCall Poisonings that were confirmed (either positive or negative) was used to evaluate the effect of Trace Poisoning nucleotide matches on the confirmation of gene identity of the cDNA fragments. These GeneCall Poisonings were then classified into four categories based on the Trace Poisoning nucleotide match (between the cDNA fragment and the associated gene sequence used in GeneCall Poisoning), such as 1/1, 2/2, 3/3, and 4/4 matches. The proportion of the positively confirmed GeneCall Poisonings to the total GeneCall Poisonings (true positives) steadily increased from about 32% in 1/1 matches to more than 80% in 4/4 matches. Similarly, the true positive rates of associating cDNA fragments to known genes was more than 80% when

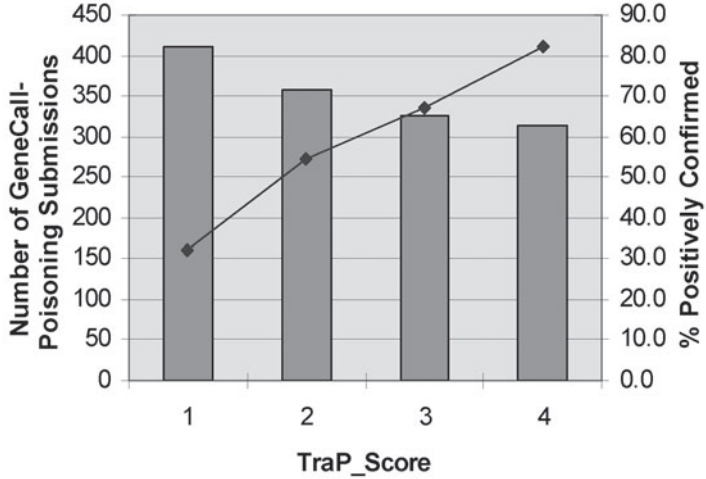


Fig. 4. The effect of Trace Poisoning on associating individual cDNA fragments with known gene sequences. The X axis represents the Trace Poisoning nucleotide matches (TraP_Score) between the cDNA fragments and the sequences in the databases. The Y1 axis shows the total number of GeneCall Poisonings submitted and the Y2 axis shows the proportion of the GeneCall Poisonings that were positively confirmed (true positives).

the size of the cDNA fragment was matched to the SeqCalling clone within adjacent 0.2 bp (data not shown).

5. Rather than up-front sequencing a large number of clones, the cDNA fragments that have a desirable expression profile but are not significantly associated with known gene sequences can be specifically queried with the SeqCalling cloned cDNA fragment library, and the corresponding clones can be queued for sequencing. This targeted sequencing approach will avoid the up-front sequencing of hundreds of thousands of clones to discover a few hundred sequences that are associated with the phenotype of interest.
6. The additional attributes from Trace Poisoning and SeqCalling may be enough to associate cDNA fragments to genes with significance values. Only a very small subset of the modulated gene fragments that are of interest in an experiment will be confirmed using GeneCall Poisoning, typically before they are used for further follow-up analysis of the cause-and-effect relationship between the gene and the phenotype of interest.

References

1. Eikhom, T. S., Abraham, K. A., and Dowben, R. M. (1975) Ribosomal RNA metabolism in synchronized plasmacytoma cells. *Exp. Cell Res.* **91,** 301–309.
2. Green, C. D., Simons, J. F., Taillon, B. E., and Lewin, D. A. (2001) Open systems: panoramic views of gene expression. *J. Immunol. Methods* **250,** 67–79.

3. Rothberg, J. M., Deem, M. W., and Simpson, J. W. (1999) Method and apparatus for identifying, classifying, or quantifying DNA sequences in a sample without sequencing. *U.S.A. Patent 5,871,697*.
4. Bruce, W., Folkerts, O., Gamaat, C., Crasta, O., Roth, B., and Bowen, B. (2000) Expression profiling of the maize flavonoid pathway genes controlled by estradiol-inducible transcription factors CRC and P. *Plant Cell* **12,** 65–80.
5. Mysore, K., Crasta, O. R., Tuori, R. P., Folkerts, O., Swirsky, P. B., and Martin, G. B. (2002) Comprehensive transcript profiling of Pto- and Prf-mediated host defense responses to infection by *Pseudomonas syringae* pv. tomato. *Plant J.* **32**, 299–315.
6. Duvick, J., Gilliam, J., Maddox, J., Crasta, O., and Folkerts, O. (2001) Amino polyol amine oxidase polynucleotides and related polypeptides and methods of use. *U.S.A. Patent 6,211,435*.
7. Duvick, J., Maddox, J., Gilliam, J., Folkerts, O., and Crasta, O. R. (2002) Compositions and methods for fumonisin detoxification. *U.S.A. Patent 6,388,171*.
8. Garnaat, C., Crasta, O., Li, Y., et al. (2000) An auxin inducible promoter for conditional complementation of male-sterile maize. Poster presented at the symposium, "Frontiers in Sexual Plant Reproduction." October 28–29, 2000, University at Albany, SUNY, Albany, NY.
9 Duvick, J., Simmons, C., Crasta, O., and Folkerts, O. (2002) Use of beta-glucosidase to enhance disease resistance and resistance to insects in crop plants. *U.S.A. Patent 6,433,249*.
10. Bruce, W., Desbons, P., Crasta, O., and Folkerts, O. (2001) Gene expression profiling of two related maize inbred lines with contrasting root-lodging traits. *J. Exp. Bot*. **52,** 459–468.
11. Kollipara, K., Saab, I. N., Wych, R. D., Lauer, M. J., and Singletary, G. W. (2002) Expression profiling of reciprocal maize hybrids divergent for cold germination and desiccation. *Plant Physiol.* **129,** 974–992.
12. Simmons, C. R., Grant, S., Altier, D. J., et al. (2001) Maize rhm1 resistance to *Bipolaris maydis* is associated with few differences in pathogenesis-related proteins and global mRNA profiles. *Mol. Plant-Microbe Inter.* **14,** 947–954.
13. Shimkets, R. A., Lowe, D. G., Tai, J. T., et al. (1999) Gene expression analysis by transcript profiling coupled to a gene database query. *Nat. Biotechnol.* **17,** 798–803.
14. Bader, J., Gold, S., Gusev, U., et al. (2002) Method of analyzing a nucleic acid. *U.S.A. Patent Application 20020015951*.
15. Gould-Rothberg, B., Ramesh, T. M., and Burgess, C. E. (2000) Integrating expression-based drug response and SNP-based pharmacogenetic strategies into a single comprehensive pharmacogenomics program. *Drug Dev. Res.* **49,** 54–64.

24

Proteomics as a Functional Genomics Tool

Ulrike Mathesius, Nijat Imin, Siria H. A. Natera, and Barry G. Rolfe

Summary

To understand the function of all the genes in an organism, one needs to know not only which genes are expressed, when, and where, but also what the protein end products are and under which conditions they accumulate in certain tissues. Proteomics aims at describing the whole protein output of the genome and complements transcriptomic and metabolomic studies. Proteomics depends on extracting, separating, visualizing, identifying, and quantifying the proteins and their interactions present in an organism or tissue at any one time. All of these stages have limitations. Therefore, it is, at present, impossible to describe the whole proteome of any organism. Plants might synthesize many thousands of proteins at one time, and the whole potentially synthesized proteome certainly exceeds the number of estimated genes for that genome. This occurs because the gene products of one gene can differ due to alternative splicing and a variety of possible posttranslational modifications. It is, therefore, essential to optimize every step towards detecting the whole proteome while realizing the limitations. We concentrate here on the most commonly used steps in high-throughput plant proteomics with the techniques we have found most reproducible and with the highest resolution and quality.

Key Words

expressed sequence tags, glycosylation, mass spectrometry, model plants, N-terminal sequencing, peptide mass fingerprinting, phosphorylation, posttranslational modifications, protein–protein interaction, proteome analysis, two-dimensional gel electrophoresis

1. Introduction

1.1. Proteomics as a Functional Genomics Tool

Biology has been revolutionized through a change from studying the function of single genes to studying the whole system of genes and their products in an organism. Genomics requires, first, the sequencing of an organism as a basis

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

for studying the existing genes and proteins and, secondly, high-throughput techniques for assaying thousands of genes, proteins, or metabolites at once ***(1,2)***.

Proteomics is the study of all the proteins produced by the genome of an organism or tissue. Compared to the genome, which is an almost unchangeable part of an organism, the transcriptome and proteome are highly variable, depending on the conditions and activities of the organism. Proteomics complements transcriptomics by providing information about the time and place of protein synthesis and accumulation, as well as identifying those proteins and their posttranslational modifications (PTMs). Gene expression does not necessarily indicate whether a protein is synthesized, how fast it is turned over, or which possible protein isoforms are synthesized. In some cases, the correlation between gene expression and protein presence is a low as 0.4 ***(3)***. First, a gene can be transcribed, but the protein is not synthesized or turned over very quickly. Second, a gene might be silent at the time, but a very stable protein might be present in the cell due to previous activity of the gene coding for it. Third, one gene can give rise to many protein products, which are the result of alternative splicing or PTM, e.g., glycosylation, phosphorylation, ubiquitinylation, and many more ***(4)***.

How does proteomics help us in the quest of plant functional genomics? Understanding gene function often starts with the identification of a mutation and description of its phenotype. But does the phenotype really result from the mutation or from downstream effects of the mutation? As gene products do not work in isolation but in a network with other gene products and metabolites, whether through protein–protein interactions or indirectly by affecting the expression of genes, gene function can only be interpreted with an understanding of all its downstream effects on the proteome and metabolome. Even in single cells, mutations in one gene often entail multiple changes in the proteome ***(5)***. In most multicellular plants, the situation is likely to be more complicated, because proteins can (inter)act differentially in different tissues. Possible cell-to-cell communication and long distance transport between plant organs of signals, mRNAs ***(6)***, or proteins means that definition of gene function always needs to include a spatial dimension in plants.

While we do not attempt to outline methods for determining protein localization or transport within the plant, we refer to the use of antibodies for *in situ* protein identification or fluorescent labeling of proteins with the green fluorescent protein or other fluorophores commonly used in plants ***(7,8)***. Protein–protein interactions can be studied with either the two-hybrid system ***(9)***, or in vivo in plants with fluorescence resonance energy transfer (FRET) ***(10)*** or fluorescence lifetime imaging (FLIM) ***(11)***, and is not detailed here.

Proteomics has been driven by recent advances in technology that enable the processing and identification of thousands of proteins in a short time. Whereas we attempt here to describe the standard protocols most well-equipped laboratories could carry out, there are a growing number of techniques for the separation and relative quantification of proteins that might be useful for high-throughput analyses for plant proteomics ***(12)***. We refer to the use of multidimensional liquid chromatography coupled with tandem mass spectrometry (MS/MS) as an alternative to two-dimensional gel electrophoresis (2-DE) for protein separation ***(13)***. Instead of quantifying protein abundance from 2-DE gels, which is an easy and standard procedure in many laboratories, a high-throughput technique, isotope-coded affinity tag-labeling (ICAT), has been developed to determine relative abundance of proteins in treated and control tissue ***(14)***.

Because proteomics is such a big task, many studies have concentrated on establishing proteomes of model plants, e.g., *Arabidopsis thaliana* ***(15)***, the model legume *Medicago truncatula* ***(16)***, rice ***(17)***, tobacco ***(18)***, and maize ***(19,20)***. The advantage of using model plants is that high-throughput MS by the commonly used method of peptide mass fingerprinting requires species-specific genomic or at least expressed sequence tag (EST) sequence information. It is more difficult to use cross-species information for protein identification, especially when proteins are posttranslationally modified.

In future, plant proteomics will involve most certainly more diverse and agronomically or pharmacologically important species, especially for understanding and exploiting secondary metabolism ***(21)*** and nodulation of plants by nitrogen-fixing bacteria ***(22,23)***. Plant proteomics will further be a useful tool for providing markers for genetic and phylogenetic analyses, and for environmental conditions, with the aim to identify candidate genes for traits like drought or pathogen resistance ***(24)***.

1.2. The Technical Background to Proteomics

1.2.1. Protein Extraction

Generally, extracting proteins from plant tissue requires tissue disruption (by grinding and sonication), separation of proteins from unwanted cell material (cell walls, water, salt, phenolics, nucleic acids) by centrifugation after precipitation of proteins with acetone–trichloroacetic acid, resolubilizing proteins in a solution that solubilizes the maximum number of different proteins, and inactivation of proteases (by acetone–trichloroacetic acid treatment or with specific protease inhibitors). Prefractionation of tissue is optional for the analysis of proteins of different organelles or microsomal fractions ***(25)***. Solubilization requires urea or, for more hydrophobic proteins, thiourea, as a chaotrope

that solubilizes, denatures, and unfolds most proteins. Nonionic zwitterdetergents, e.g., 3-[3-cholamidopropyl)-dimethyl-ammonio]-1-propane sulfonate (CHAPS), Triton®-X, or amidosulfobetaines are used to solubilize and separate proteins in the mixture ***(26)***. Sodium dodecyl sulfate (SDS) is also a strong detergent and used to solubilize membrane proteins. However, it renders a negative charge to proteins and, therefore, interferes with isoelectric focusing. Reducing agents (usually dithiothreitol [DTT], 2-mercaptoethanol, or tributyl phosphine) are needed to disrupt disulfide bonds.

1.2.2. First-Dimension Isoelectric Focusing

Isoelectric focusing separates proteins according to their isoelectric points (pIs). The use of immobilized pH gradients enables clear separation of proteins over different pH ranges ***(27,28)***. The resolution depends on a clean sample preparation (presence of salts or SDS will introduce charge and interfere with protein focusing), a slowly increasing electric field, and a high final voltage (1000–3500 V) to enable proteins to move into the gel. Proteins, each with a different pI, will move into the first dimension gel–strip that contains an immobilized pH gradient, until their net charge is zero, i.e., the pH along the gel equals their pI.

1.2.3. 2-DE

In the second dimension, proteins are separated by their molecular mass. First, proteins are treated with SDS, which introduces negative charges proportional to the size of the protein. Second, proteins are separated by an electric current through a polyacrylamide gel with a certain pore size. Therefore, this technique is referred to as SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Depending on the polyacrylamide concentration, proteins in the range from approx 5–200 kDa can be separated on a gel.

1.2.4. Protein Identification

Protein identification has been possible on small scales and with limitations by N-terminal or C-terminal sequencing. Recent improvements in MS have made it possible to identify proteins faster, on a larger scale, with less protein amounts. In addition, posttranslational modifications can be determined by MS/MS analysis, and proteins can be identified even when bound to other proteins in complexes ***(29)***. A standard technique for protein identification with matrix-assisted laser desorption–ionization time of flight (MALDI-TOF) MS is peptide mass fingerprinting ***(30,31)***. Proteins are excised from a gel or eluted from a column and digested by specific proteases to generate peptides. The masses of these peptides are measured by MALDI-TOF and compared to theoretical

digests of translation products of genomic or EST databases of the organism. While this technique requires sequence information, it is fast, specific, and allows limited identification of PTMs by detecting shifts in the masses of predicted proteins by the specific mass of a modification, e.g., phosphorylation or ubiquitinylation. Detection of PTMs is necessary, especially for phospho- or glycoproteins, because they affect protein function. Phosphorylation can be detected by the use of antiphosphotyrosine antibodies on blots of 2-DE ***(32)***, similar to a Western blot, or by radiolabeling of proteins and detecting the labeled proteins ***(33)***. Both techniques have been extensively detailed elsewhere and are not described here. Glycosylation of proteins can easily be detected on gels by periodic acid Schiff reaction ***(34)***. In addition, specific enzymes can be used for selective cleavage of several common PTMs ***(35)***. More sophisticated identification of specific glycosylations require MS/MS analysis and will be beyond the scope of this chapter. For a summary of standard proteome analysis procedures, *see* **Fig. 1**.

In summary, plant proteomic studies should carefully consider the following points before starting their analyses:

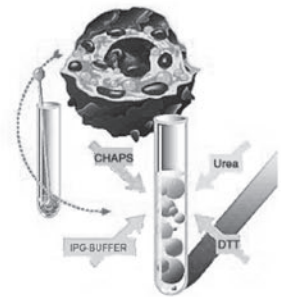
1. Using a species for which genomic sequence is available is the easiest approach, and protein identification by peptide mass fingerprinting (*see* **Subheading 3.3.1.**) is a fast and high-throughput approach.
2. If working with an unsequenced species, N-terminal sequencing, a slower and less sensitive technique, or tandem MS may be necessary.
3. With any species, it is essential to be precise about tissue and cell type separation before protein extraction to gain spatial information about protein accumulation and to reproduce the results that are highly dependent on tissue type and growth conditions. Good knowledge about the biological material, its quality, and reproducibility are the most important prerequisites for successful proteome analysis.
4. Keep in mind that it is very difficult at present to see the whole proteome of any one organism, especially low abundance, very basic, and integral membrane proteins.

2. Materials

2.1. Protein Extraction from Plant Material

2.1.1. Extraction of Soluble Proteins

Before tissue harvesting, prepare the following solutions. All solutions should be made fresh. Sample buffer can be frozen at –80°C for several months, however, the protease inhibitors phenylmethylsulfonyl fluoride (PMSF) and ethylenediaminetetraacetic acid (EDTA) disodium salt have to be added fresh before use. Never heat the frozen buffer after addition to the proteins to more than 37°C, because the urea can modify the protein charge by carbamylation.

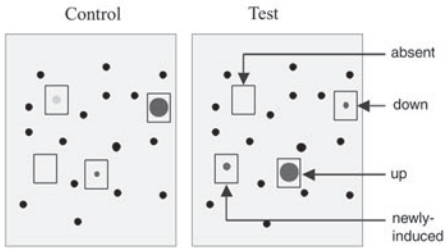


solubilisation

reduction/
alkylation

Sample preparation
- prefractionation
- sequential extraction
- solubilisation
- removal of interfering compounds

Protein separation by IEF
- narrow range
- broad range
- basic



excision of spots
from gels/blots

tryptic digestion

peptide mass
measurement by MALDI

Image analysis
- differential display
- quantitation
- experimental pI and MW
- bioinformatics

Pre-processing prior to analysis
- excision of spots
- destaining
- digestion of proteins with proteolytic enzymes
- peptide extraction
- peptide concentration
- enrichment of specific peptides such as posphopeptides

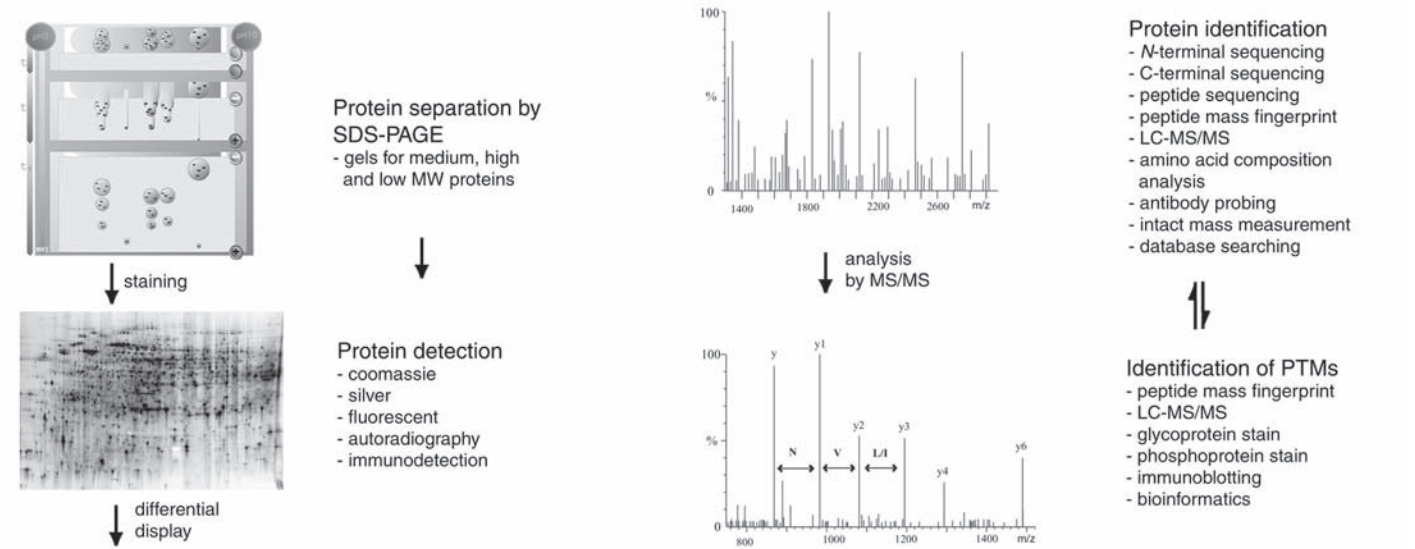


Fig. 1. Flow chart for the analysis of plant proteomes by 2-DE followed by MS protein identification. Proteins are extracted, solubilized and separated by 2-DE, after which they are visualized and differentially displayed and selected for identification. The protein spots are excised, SDS, stain and salts are removed, and the proteins are digested with a site-specific protease, usually trypsin. Peptides are subsequently extracted, and a fraction is used to determine the mono-isotopic peptide ion masses by MALDI-TOF MS. The experimental peptide ion masses are then searched against proteins predicted from genomic or EST sequence data. If no positive identification can be made, (partial) sequences of the remaining peptides are determined by MS/MS. The sequence tags are used to search against EST or annotated genomic sequence data for protein identification. Posttranslational modifications can be detected by MALDI-MS, MS/MS, and/or in combination with glyco- or phosphoprotein staining kits or by immunoblotting.

1. TCA/acetone: 10% trichloroacetic acid in acetone with 0.07% (0.45 m*M*) DTT or 0.07% (9 m*M*) 2-mercaptoethanol on dry ice.
2. Acetone with 0.07% DTT on dry ice (approx 3 mL/6 mL for each gram of fresh tissue weight for **steps 1** and **2**, respectively).
3. Sample buffer containing: 9 *M* urea, 4% (w/v) CHAPS, 1% (w/v) DTT, 1% (v/v) BioLyte 3–10 ampholytes (Bio-Rad), 35 m*M* Tris base, 1 m*M* PMSF, and 5 m*M* EDTA. Keep at room temperature (20°C).

2.1.2. Extraction and Solubilization of Plant Membrane Proteins

Make all solutions fresh before use.

1. Membrane extraction buffer: 0.5 m*M* (N-[2 hydroxyethyl]piperazine-N'-[2-ethanesulfonic acid] (HEPES)/KOH, pH 7.8, 0.5 m*M* sucrose, 10 m*M* PMSF, 5 m*M* DTT, 1 m*M* ascorbic acid, and 0.6% (polyvinylpyrrolidone [PVP], molecular weight [MW = 40,000], to remove interfering phenolic compounds).
2. 4% SDS solubilization buffer: 100 m*M* Tris-HCl, pH 6.8, 4% SDS, 20% glycerol, and 5 m*M* DTT.
3. 0.5% SDS solubilization buffer: 100 m*M* Tris-HCl, pH 6.8, 0.5% SDS, and 5 m*M* DTT.
4. Urea–thiourea solubilization buffer: 50 m*M* Tris-HCl, pH 6.8, 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 5 m*M* DTT, 0.5% Triton X-100.

2.2. 2-DE

2.2.1. Isoelectric Focusing

One day before isoelectric focusing, prepare rehydration solution containing 8 M urea, 0.5% (w/v) CHAPS, 0.15% (w/v) DTT, 0.5% (v/v) Biolyte 3–10 or 6–11 ampholytes, and a trace of bromophenol blue. You will need 360 µL for an 18-cm strip and 480 µL for a 24-cm strip. When separating proteins on basic pH gradients, addition of 10% isopropanol to the rehydration solution can help to avoid streaking due to electroendo-osmosis along the strip ***(36)***.

2.2.2. SDS-PAGE

1. Equilibration solution (1): 40% (v/v) glycerol, 0.05 *M* Tris-HCl, pH 6.8, 6 *M* urea, 2% (w/v) SDS, and 2% (w/v) DTT.
2. Equilibration solution (2): 40% glycerol, 0.05 *M* Tris-HCl, pH 6.8, 6 *M* urea, 2% (w/v) SDS, 2% (w/v) iodoacetamide (to prevent protein oxidation), and 0.005% (w/v) bromophenol blue.
3. Prepare molecular weight markers according to the specific manufacturer's protocol.

2.2.3. Protein Staining

2.2.3.1. Silver Staining

Prepare all solutions fresh before use and with highest grade chemicals, all in ultrapure water.

1. Fixation solution: 10% (v/v) acetic acid, 40% (v/v) ethanol, 50% ultrapure water.
2. Sensitizer: 30% (v/v) ethanol, 4.1% (w/v) sodium acetate, 0.275% (w/v) potassium tetrathionate, and 0.5% (v/v) glutaraldehyde.
3. Silver stain: 0.2% (w/v) silver nitrate, 0.062% (w/v) HEPES, 0.07% (v/v) formaldehyde.
4. Developer: 3% (w/v) potassium carbonate, 0.0012% (w/v) sodium thiosulfate, 0.025% (v/v) formaldehyde.
5. Stop solution: 5% (w/v) Tris-base, 2% (v/v) acetic acid.

2.2.3.2. Coomassie® Staining

1. Coomassie staining solution: 10% (w/v) ammonium sulfate, 2% (w/v) phosphoric acid, 0.1% (w/v) Coomassie Brilliant Blue G250 (Bio-Rad).
2. Stop solution: 0.1 *M* tris-phosphoric acid, pH 6.5.
3. Destain: 25% methanol.
4. Fixation solution: 20% ammonium sulfate.

2.2.3.3. SYRPO® Ruby Staining

Fixation solution: 7% acetic acid and 10% methanol.
Staining solution: dilute premade SYPRO solution as specified by each manufacturer.

2.3. Protein Identification

2.3.1. Peptide Mass Fingerprinting and Tandem MS Analysis

1. Destain: 100 m*M* ammonium bicarbonate, pH 7.8.
2. Trypsin solution: 8 μL of 15 ng/mL sequencing-grade modified trypsin (Promega) in 25 m*M* ammonium bicarbonate, pH 7.8.
3. 50% (v/v) Acetonitrile and 0.5% (v/v) trifluoroacetic acid.
4. Matrix: α-cyano-4-hydroxcinnamic acid, 10 mg/mL in 70% (v/v) acetonitrile, 1% (v/v) trifluoroacetic acid.

2.3.2. Detection of Glycoproteins in Gels

1. Fixation solution: 50% methanol.
2. Washing solution: 3% acetic acid.
3. Oxidization solution: 1% periodic acid, 3% acetic acid in distilled water.
4. Glycoprotein staining solution: dilute the stock Pro-Q Emerald 300 dye solution 1:50 in Pro-Q Emerald 300 dilution buffer (both from Molecular Probes) just prior to staining.

3. Methods

3.1. Protein Extraction from Plant Material

3.1.1. Extraction of Soluble Proteins

1. Grind tissue in liquid nitrogen after addition of fine glass powder (0.01–0.1 mm grain size; Schott; add approx 10% of the tissue vol) in a mortar and pestle and suspend in –20°C cold TCA/acetone in an acetone-resistant centrifuge tube on dry ice. It is important to grind tissue thoroughly in liquid nitrogen, the finer the powder, the better the yield. For disruption of soft plant material, especially when isolating organelles, tissue can be ruptured by vortex mixing or osmotic lysis, however, most plant tissue will not sufficiently rupture until ground in liquid nitrogen. The yield should range from 0.5–2 mg of protein from every gram of tissue fresh weight, but might change with tissue type, depending on water, fiber, and phenolics content.
2. Sonicate the suspension on dry ice with a probe sonicator at approx 20 MHz 6× for 10 s each with intermittent 1 min breaks to avoid overheating.
3. Leave samples for 1 h at –20°C.
4. Centrifuge samples at 35,000*g* for 15 min at 4°C.
5. Discard the supernatant and resuspend the pellet in cold (–20°C) acetone containing 0.07% (w/v) DTT.
6. Place samples at –20°C for 30 min and then centrifuge at 12,000*g* for 15 min at 4°C.
7. Repeat the last washing step.
8. Briefly lyophilize the pellet (3–5 min) to evaporate any acetone, and suspend the dry pellet in sample buffer (approx 500 mL for every gram fresh weight) by sonication in ice cold water in a sonic bath (1–3 min) and vortex mixing for several minutes.
9. Centrifuge sample at 12,000*g* for 15 min at 20°C and collect the supernatant that should contain the solubilized proteins.
10. Repeat the solubilization step and pool the supernatants.
11. Measure the protein concentration of the sample (e.g., with a Bradford assay) and keep at –80°C until used for isoelectric focusing (*see* **Note 1**).

3.1.2. Extraction and Solubilization of Plant Membrane Proteins (see Note 2)

1. Grind plant tissues (10 g) under liquid nitrogen, add 20 mL of membrane extraction buffer, and filter through two layers of miracloth (Calbiochem-Novabiochem) while keeping the suspension below 4°C.
2. Centrifuge the suspension for 15 min at 15,000*g* at 4°C.
3. Collect the supernatant and centrifuge again for 35 min at 105,000*g* at 4°C. Collect the remaining pellets.
4. Solubilize the membrane proteins in 2 mL of 4% SDS solubilization buffer by boiling for 5 min.

5. Add a 10-fold vol of prechilled acetone to the solution and incubate for at least 1 h at –20°C.
6. After centrifugation for 5 min at 12,000*g* at 4°C, lyophilize the remaining pellet and solubilize in 200 μL of 0.5% SDS solubilization buffer for 5 min.
7. Dilute the solution with 750 μL urea–thiourea solubilization buffer and solubilize by sonication as described in **Subheading 3.1.1.** and then centrifuge for 5 min at 12,000*g* at room temperature.
8. Protein concentration can be determined in SDS buffer using the Lowry's protein assay ***(37)*** as the Bradford assay is not compatible with this method. SDS must be diluted as at least eight-fold with the urea solubilization buffer as it interferes with isoelectric focusing.

3.2. 2-DE

3.2.1. Isoelectric Focusing

Depending on the electrophoresis system used, variations of this protocol can be used. We describe here the use of a Multiphor II horizontal electrophoresis system (Amersham Pharmacia Biotech, Piscataway, NJ, USA) for isoelectric focusing and SDS-PAGE because of its high reproducibility and resolution ***(27)***. Immobilized pH gradient (IPG) strips are highly recommended. Samples can be cup-loaded at the anode or cathode or can be rehydrated into the strip in place of the rehydration solution (*see* **Note 3**). For loading an 18–24 cm strip, use approx 150–200 μg of protein for analytical (silver- and SYPRO Ruby-stained) and 800–1000 μg for preparative (Coomassie stained) gels. Isoelectric focusing markers can be loaded together with the sample to enable accurate determination of pI values of the sample proteins. Make sure to cover strips with paraffin oil to avoid drying out and crystallizing of the sample. Focus the rehydrated strips at 20°C, 1 mA and 5 W for a total of 200 kVh with the following voltage gradients: 30 min at 150 V, 5 min on a linear gradient from 150–300 V, 6 h at 300 V, 5 h on a linear gradient from 300–3500 V and 54 h at 3500 V (*see* **Note 4**). After isoelectric focusing, strips can either be wrapped in plastic foil and frozen at –80°C or directly equilibrated for SDS-PAGE.

3.2.2. SDS-PAGE

1. Equilibrate the IPG strips for 10 min in equilibration solution 1 and for another 10 min in equilibration solution 2.
2. Gently blot strips dry on filter paper to remove excess equilibration solution without damaging the gel.
3. Load the IPG strip onto the second dimension gel together with molecular weight markers at one or both ends of the strip. Ensure good contact with the gel and avoid air bubbles. Also avoid water drops on horizontal gels as these will distort

the separation pattern. Usually, electrophoresis is carried out at 6°–15°C (*see* **Note 5**) at 300 V for 1 h and subsequently at 600 V for 4 to 5 h, until the bromophenol blue front reaches the end of the gel. These conditions will depend on the apparatus used for SDS-PAGE.

3.2.3. Protein Staining

3.2.3.1. Silver Staining

For successful silver staining, use only high purity fresh chemicals. Especially formaldehyde and glutaraldehyde solutions should be made fresh. We never use stock solutions but make up all solutions just before use. To reduce impurities, all solutions can be vacuum-filtered through 0.45-µm membranes just before use.

Staining is easiest done in photographic trays on an orbital shaker (50 rotations/min) at room temperature under a fume hood.

1. Directly after SDS-PAGE, fix gels 3×, for 30 min each, in fixative.
2. Change to sensitizer for 16 h, avoid evaporation by covering the tray carefully with plastic wrap.
3. Wash the gel at least 6× in ultrapure water for 30 min each (more washing is better than less).
4. Incubate gels with silver staining solution for 2 h in the dark. To wash of the silver, quickly rinse for 10 s in ultrapure water. If this washing step is done for any longer, the silver will disassociate from the protein spots.
5. Develop gels for 5–7 min and stop the development by replacing the developer with stop solution. The development should be stopped when no more new spots become visible and before the background of the gel becomes dark (*see* **Note 6**). Timing is important because gels of different runs need to be comparable in staining.
6. Leave the gels in stop solution for no longer than 20 min to avoid color changes in the silver stain.
7. Wash gels in distilled water several times, scan, and store sealed in plastic pouches in a few milliliters of 1% methanol to discourage microbial contamination. Gels can be stored at room temperature or at 4°C for many years.

3.2.3.2. Coomassie Staining

1. Directly after SDS-PAGE, rinse gels in ultrapure water for 1 min (optional) and stain with 100 mL Coomassie staining solution for 20 h in a sealed plastic pouch or tray (prevent evaporation).
2. Wash gel in a tray for 3 min in 0.1 *M* Tris-phosphoric acid, and destain for 1 min in 25% methanol and fix the proteins with a 24-h wash in 20% ammonium sulfate. This staining procedure can be repeated 3× in total (*see* **Note 7**).
3. If the gel is to be used for protein identification, scan the gel, then cut out proteins

from the gel as soon as possible, before contamination or chemical modification of the proteins can occur.
4. Store gel in purified, sterile water containing 1% methanol in a sealed pouch at 4°C or room temperature.

3.2.3.3. SYPRO Ruby Staining

SYPRO Ruby fluorescent staining is an easy and a highly sensitive (detection limit of 1 ng/spot) staining with broad linear quantitation range. Use polyvinyl chloride photographic staining trays. Glass dishes are not recommended.

1. Incubate gels (245 × 180 × 0.5 mm) in a fixative solution for 30 min to 1 h and then in 300 mL of SYPRO Ruby Protein Gel stain (Molecular Probes) for 3 h to overnight. It is critical that the stain vol used at least 10× the vol of the gel. Perform all staining and washing steps with continuous gentle agitation (e.g., on an orbital shaker at 50 rpm).
2. After staining, rinse the gels in deionized water for 30–60 min to wash residual dye out of the polyacrylamide matrix. SYPRO Ruby gel stain has two excitation maxima (approx 280 and 450 nm) and has an emission maximum near 618 nm. Proteins stained with the dye can be visualized using a 300-nm UV transilluminator, a blue-light transilluminator, or a laser scanner. Use new UV lamps for visualization as old ones limits detection sensitivity. Accurate quantitation can be achieved using a charge-coupled device (CCD) camera or a laser scanner. Remove plastic backing of the SDS gels before scanning with laser scanner as the plastic backing interferes with the dye.

3.3. Protein Identification

3.3.1. Preparation for MALDI-TOF MS

1. Excise protein spots manually with a sterile scalpel blade (one per protein) from Coomassie- or SYPRO Ruby-stained gels and store in a drop of 50% high-grade methanol. To avoid or reduce contamination with human keratin, one should wear gloves and preferably work in a laminar-flow hood.
2. Destain each spot and spin-dry with a speed vac for several minutes.
3. Digest proteins in gel with trypsin solution for 16 h at 37°C.
4. After the tryptic (or other protease) digestion, extract the peptides with 50% acetonitrile and 0.5% trifluoroacetic acid.
5. Spot a 1-µL aliquot onto a sample plate along with 1 µL of matrix and air-dry.
6. Submit each digested protein for MALDI-TOF MS (*see* **Note 8**) to determine the masses of the tryptic fragments. These can be obtained in the form of a list of peptide fragments or as a spectrum for visual inspection.

3.3.2. Peptide Mass Fingerprinting

1. Next, the generated spectra have to be compared against theoretical spectra. For this, Internet-available sites can be used free of charge for searches against com-

mon organisms, e.g., MOWSE (http://www.hgmp.mrc.ac.uk/Bioinformatics/Webapp/mowse/), Mascot (http://www.matrixscience.com/), PepIdent (http://au.expasy.org/tools/), or Profoud (http://prowl.rockefeller.edu/).
2. Alternatively, create your own database of your specific organism for more specific searches using either genomic or EST sequences. Genomic sequence is preferable and can be obtained for many organisms, e.g., from the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). EST sequences can be used as well, but will produce less reliable results, because they usually do not cover the whole coding sequence and hence, will not generate the full expected spectrum of a theoretical tryptic digest.
3. Download an organism's genome or EST sequence database in FASTA format.
4. To generate translation products, e.g., use the FLIP program developed in the Sequencing Unit of the Organelle Genome Megasequencing Program (OGMP), (http://megasun.bch.umontreal.ca/ogmp/ogmpid.html). Homology comparisons can be done more specifically and successfully with software packages like MassLynx (Micromass).
5. The trypsin autodigestion peaks at 842.51 and 2211.1 Da are usually used for internal calibration.
6. Stringency of matching is paramount to avoid selecting the wrong matches. Generally, a minimum of four peptide matches, a maximum of one miscleavage per peptide, >20% sequence coverage, and 100 ppm molecular weight discrepancy are acceptable.

3.3.3. MS/MS Analysis

Excise, destain, and digest selected spots as described in **Subheading 3.3.1.** and submit each digested protein for MS/MS (*see* **Note 9**) to obtain sequence information or to elucidate possible posttranslational modifications. Search engines, such as ProteinProspector (http://prospector.ucsf.edu/) and Mascot (http://www.matrixscience.com/), can be used for protein identification.

3.3.4. Detection of Glycoproteins in Gels

Perform all staining and washing steps with continuous, gentle agitation (e.g., on an orbital shaker at 50 rpm).

1. Fix gels in fixation solution for 30–45 min and wash 2× in washing solution for 5–10 min each.
2. Oxidize the carbohydrates by incubation in oxidization solution for 20–30 min.
3. Wash the gel 4× for 5–10 min each in washing solution to remove the residual periodate.
4. Incubate the gel in 250 mL glycoprotein staining solution for 30–120 min. The diluted reagent degrades upon long-term storage, so only the amount required for staining should be prepared.
5. Incubate the gel in washing solution 2× for 5–10 min each.

6. Pro-Q Emerald dye has an excitation maximum at approx 280 nm and an emission maximum near 530 nm. Proteins stained with the dye can be visualized using a 300-nm UV transilluminator, a blue-light transilluminator or a laser scanner. Accurate quantitation can be achieved using a CCD camera or a laser scanner. Remove plastic backing of the SDS gels before staining as the plastic backing interferes with the dye. Gels can be stored in washing solution.

3.4. Quantification of Protein Abundance

3.4.1. Gel Scanning and Image Analysis

Scan silver-stained gels on a high resolution scanner equipped with transparency adaptor and Coomassie-stained gels with an opaque white background in reflective mode at a minimum of 600 dots per square inch (dpi) and save them as TIF images for subsequent analysis. Several software packages are available for spot quantitation (*see* **Note 10**). These software packages allow quantification of spot vol, areas, and other parameters. For quantification of protein abundance, use a percentage vol as the relative abundance of a spot compared to the total proteins displayed on the gel. This will take into account variations between differently stained gels. In addition, it is useful to load protein markers of known size, pI, and amount and use these for calibration of spot position and protein abundance. Manual annotation of protein spots and careful cross-checking of protein identity on gels that are to be compared is required to ensure that the areas of each protein are correctly assigned and identified.

3.4.2. Statistical Analysis and Reproducibility

After quantifying protein spots across repeat gels and between treatments, the significance of the observed changes in protein abundance can be calculated with analysis of variance (*see* **Note 11**).

4. Notes

1. The suggested extraction protocol has been successfully used for various species and tissues. To maximize extraction of hydrophobic proteins from tissue, it is possible to re-extract the insoluble pellet of the last step with organic solvents, e.g., chlorophorm–methanol. This can extract some more insoluble proteins, however, we did not find any great improvement over the standard extraction buffer, which is effective at solubilizing peripheral membrane proteins. Protein yields should be approx 0.5–2 mg protein/g fresh tissue weight. If protein concentrations are too low, or vol too large to load onto the first dimension, either precipitate proteins again with TCA/acetone and resolubilize in a smaller vol of sample buffer, or reduce the vol with a Centricon® (Amicon®) column (Millipore).
2. Extraction of membrane proteins is inherently difficult, and it is so far not possible to extract all membrane proteins, especially integral membrane proteins. A good overview on membrane protein isolation and 2-DE can be found in refs. ***(38,39)***.

3. We found best results with cup-loading samples at the anode. This results in sharper focusing than in gel rehydration. In gel rehydration might be necessary when large sample vol have to be loaded, e.g., for Coomassie-stained gels. In this case, replace the rehydration solution of the IPG strip with the sample itself, after adding a trace of bromophenol blue.
4. In our experience, isoelectric focusing improves if focusing times are increased from the usually recommended 25–30 kVh (for 18-cm strips) to about 200 kVh without leading to overfocusing. Focusing times will need to be adjusted for strips from other manufacturers than those mentioned here as specified. Problems with isoelectric focusing are often unsharp spots, resulting in horizontal streaks across the 2-DE gel, especially in the alkaline pH range. This can be improved by adding 10% (v/v) glycerol and 10% (v/v) isopropanol to the rehydration solution. Salt content in the sample can also interfere with focusing, and if salt is present, it should be removed from the sample before isoelectric focusing.
5. The temperatures used for SDS-PAGE vary and might depend on the sample. We found good results with running horizontal gels at 6°C. However, to increase protein transfer of membrane proteins from the IPG strip to the gel, the temperature may be increased to 15°C for the first 30 min or longer. In any case, an efficient cooling plate is needed to avoid temperature differences between gel edges and center, otherwise "smiling" of the gel front might occur. The choice of uniform vs gradient 2-DE gels and their composition depends on the proteins of interest. A gradient gel of 12–14% resolves most proteins between 6–150 kDa. Smaller size proteins can be resolved on higher percentage gels (e.g., 15–25%), larger proteins will resolve better on low percentage (e.g., 7–10% gels). Gradient gels have the advantage of resolving both smaller and larger proteins, but at the same time, the whole gel image will be compressed into a smaller area, resulting in lower resolving power for large numbers of proteins.
6. When silver staining results in only faintly stained spots, the formaldehyde solution might be old. In this case, resensitize and restain the gel in silver solution with a new or a larger vol of formaldehyde.
7. Timing of Coomassie staining is not as critical as for silver staining. The Coomassie staining solution, which should not be reused more than twice, can be left for longer than indicated.
8. PMF can be done on a Micromass TofSpec 2E Time of Flight Mass Spectrometer (Waters).
9. Tandem MS can be used to obtain sequence information of the peptides generated following enzymatic digestion of an individual isolated protein. Tandem MS analysis can be done on a Quadrupole TOF system or on an Ion Trap system.
10. Software packages for 2-DE image analysis and quantification include Melanie (Swiss Institute of Bioinformatics), PDQuest (Bio-Rad), Imagemaster (Amersham Pharmacia Biotech), Progenesis (Nonlinear Dynamics) or Z3 (Compugen).
11. Use a minimum of three repeats for every sample to evaluate variations between gel runs. The horizontal, precast gels should produce a high reproducibility of gels (>95% of protein spots present in each gel in the same relative gel positions).

If variations between gels of the same biological material are large, spot quantitation will not produce meaningful results.

References

1. Godovac-Zimmermann, J. and Brown, L. R. (2001) Perspectives for mass spectrometry and functional proteomics. *Mass Spectrom. Rev*. **20,** 1–57.
2. Rouquie, D., Peltier, J. B., Marquismansion, M., et al. (eds.) (1997) *Proteome Research: New Frontiers in Functional Genomics*. Springer, Berlin.
3. Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18,** 533–537.
4. Battey, N. H., Dickinson, H. G., and Hetherington, A. M. (eds.) (2001) *Post-Translational Modifications in Plants*. Cambridge University Press, Cambridge.
5. Guerreiro, N., Ksenzenko, V. N., Djordjevic, M. A., Ivashina, T. V., and Rolfe, B. G. (2000) Elevated levels of synthesis of over 20 proteins results after mutation of the *Rhizobium leguminosarum* exopolysaccharide synthesis gene *pssA*. *J. Bacteriol*. **182,** 4521–4532.
6. Jorgensen, R. A., Atkinson, R. G., Forster, R. L. S., and Lucas, W. J. (1998) An RNA-based information superhighway in plants. *Science* **279,** 1486–1487.
7. Mason, W. T. (ed.) (1999) *Fluorescent and Luminescent Probes for Biological Activity. A Practical Guide to Technology for Quantitative Real-Time Analysis*. Academic Press, London.
8. Wouters, F. S., Verveer, P. J., and Bastiaens, P. I. H. (2001) Imaging biochemistry inside cells. *Trends Cell Biol*. **11,** 203–211.
9. Legrain, P., Wojcik, J., and Gauthier, J. M. (2001) Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet*. **17,** 346–352.
10. Gadella, T. W. J., van der Krogt, G. N. M. and Bisseling, T. (1999) GFP based FRET microscopy in living cells. *Trends Cell Biol*. **4**, 287–291.
11. Bastiaens, P. I. H. and Squire, A. (1999) Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. *Trends Cell Biol*. **9,** 48–52.
12. Van Wijk, K. J. (2001) Challenges and prospects of plant proteomics. *Plant Physiol.* **126,** 501–508.
13. Link, A. J., Eng, J., Schieltz, D., et al. (1999) Direct analysis of protein complexes by mass spectrometry. *Nat. Biotechnol*. **17,** 676–682.
14. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol*. **17,** 994–999.
15. Kamo, M., Kawakami, T., Miyatake, N., and Tsugita, A. (1995) Separation and characterization of *Arabidopsis thaliana* proteins by two-dimensional gel electrophoresis. *Electrophoresis* **16,** 423–430.
16. Mathesius, U., Keijzers, G., Natera, S. H. A., Weinman, J. J., Djordjevic, M. A., and Rolfe, B. G. (2001) Establishment of a root proteome reference map for the model legume *Medicago truncatula* using the expressed sequence tag database for peptide mass fingerprinting. *Proteomics* **1,** 1424–2440.

17. Imin, N., Kerim, T., Weinman, J. J., and Rolfe, B. G. (2001) Characterisation of rice anther proteins expressed at the young microspore stage. *Proteomics* **1,** 1149–1161.
18. Rossignol, M. (1997) Construction of a directory of tobacco plasma membrane proteins by combined two-dimensional gel electrophoresis and protein sequencing. *Electrophoresis* **18,** 654–660.
19. Porubleva, L., Vander Velden, K., Kothari, S., Oliver, D. J., and Chitnis, P. R. (2001) The proteome of maize leaves: use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints. *Electrophoresis* **22,** 1724–1738.
20. Touzet, P., Riccardi, F., Morin, C., et al. (1996) The maize two dimensional gel protein database—towards an integrated genome analysis program. *Theor. Appl. Genet*. **93,** 997–1005.
21. Jacobs, D. I., van der Heijden, R., and Verpoorte, R. (2000) Proteomics in plant biotechnology and secondary metabolism research. *Phytochem. Anal.* **11,** 277–287.
22. Natera, S. H. A., Guerreiro N., and Djordjevic, M. A. (2000) Proteome analysis of differentially displayed proteins as a tool for the investigation of symbiosis. *Mol. Plant-Microbe Interact*. **13,** 995–1009.
23. Morris, A. C. and Djordjevic, M. A. (2001) Proteome analysis of cultivar-specific interactions between *Rhizobium leguminosarum* biovar trifolii and subterranean clover cultivar Woogenellup. *Electrophoresis* **22,** 586–598.
24. Thiellement, H., Bahrman, N., Damerval, C., et al. (1999) Proteomics for genetic and physiological studies in plants. *Electrophoresis* **20,** 2013–2026.
25. Jung, E., Heller, M., Sanchez, J.-C., and Hochstrasser, D. F. (2000) Proteomics meets cell biology: the establishment of subcellular proteomes. *Electrophoresis* **21**, 3369–3377.
26. Chevallet, M., Santoni, V., Poinas, A., et al. (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* **19,** 1901–1909.
27. Görg, A., Boguth, G., Obermaier, C., Posch, A., and Weiss, W. (1995) Two-dimensional polyacrylamide gel electrophoresis with immobilized pH gradients in the first dimension (IPG-dalt)—the state of the art and the controversy of vertical versus horizontal systems. *Electrophoresis* **16,** 1079–1086.
28. Görg, A., Obermaier, C., Boguth, G., et al. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21,** 1037–1053.
29. Gevaert, K. and Vandekerckhove, J. (2000) Protein identification methods in proteomics. *Electrophoresis* **21,** 1145–1154.
30. Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of protein by peptide mass fingerprinting. *Curr. Biol*. **3,** 327–332.
31. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90,** 5011–5016.

32. Stancato, L. F. and Petricoin, E. F., III. (2001) Fingerprinting of signal transduction pathways using a combination of anti-phosphotyrosine immunoprecipitations and two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **22,** 2120–2124.
33. Yan, J. X., Packer, N. H., Gooley, A. A., and Williams, K. L. (1998) Protein phosphorylation—technologies for the identification of phosphoamino acids. *J. Chromatog*. **808,** 23–41.
34. Steinberg, T. H., Pretty On Top, K., Berggren, K. N., et al. (2001). Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **1,** 841–855.
35. Tsugita, A., Kiyazaki, K., Nabetami, T., Nozawa, T., Kamo, M., and Kawakami, T. (2001) Application of chemical selective cleavage methods to analyze post-translational modifications in proteins. *Proteomics* **1,** 1082–1091.
36. Görg, A., Obermaier, C., Boguth, G., Csodas, A., Diaz, J.-J., and Madjar, J.-J. (1997) Very alkaline immobilised pH gradients for two-dimensional electrophoresis of ribosomal and nuclear proteins. *Electrophoresis* **18,** 328–337.
37. Lowry, O. H., Rosebrough, J. N., Farr, A. L., and Randall, R. J. (1951) Protein measurements with the Folin reagent. *J. Biol. Chem*. **193,** 265–275.
38. Santoni, V., Molloy, M., and Rabilloud, T. (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21,** 1054–1070.
39. Molloy, M. (2000) Two-dimensional gel electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem*. **280,** 1–10.

25

Metabolite Profiling as a Functional Genomics Tool

Anusha P. Dias, Johnny Brown, Pierluigi Bonello, and Erich Grotewold

Summary

Plants accumulate a very large number of small molecules (phytochemicals) with important functions in the ecology of plants and in the protection against biotic and abiotic stress conditions. Little is known on how phytochemical biosynthetic pathways are regulated, which is a key step to successfully engineering plant metabolism. Plant natural products are usually not essential, and genetic analyses often fail to identify phenotypes associated with the absence of these compounds. We have investigated the use of metabolite profiling of plant cells in culture to establish the function of transcription factors suspected to control plant metabolic pathways.

Key Words

natural products, transcription factors, R2R3 Myb, culture cells, GC/MS, HPLC

1. Introduction

Metabolic engineering involves the modification of biochemical networks to alter the accumulation of specific metabolites. Overexpression of the enzymes involved in rate-limiting steps has been effective in some cases to manipulate the levels of product accumulation. These studies, however, are limited due to flux considerations, metabolite channeling, and homeostatic control of metabolic pathways ***(1)***. Transcription factors are emerging as powerful tools that allow the simultaneous activation of multiple genes in a pathway, overcoming the main limitations associated with flux constrains ***(2,3)***. At present, a drawback in using transcription factors to manipulate plant metabolism is how little is known on the regulation of plant biosynthetic pathways. Classical loss-of-function mutant approaches are often unsuccessful in uncovering phenotypes associated with mutations in genes involved in the biosynthesis or regulation of metabolic pathways, due to gene redundancy and to the

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

inherent plasticity of plant metabolism. Thus, it is imperative to develop novel methods that do not rely on phenotype to establish the function of genes involved in plant metabolism.

The method described here monitors variations in the accumulation of metabolites in plant cells in culture, ectopically expressing transcription factors, as a hypothesis-generating tool to establish the possible pathways regulated by particular regulatory proteins. The first step consists of generating a transgenic cell line expressing the regulator from a constitutive or inducible promoter. The second step is to subject extracts from transformed and control cells to various metabolic profiling approaches to determine the qualitative and quantitative differences in metabolite accumulation ***(4–6)***. Numerous analytical techniques are available to monitor and purify individual metabolite ***(7)***, but for a high-throughput analysis rate, a one-by-one approach would become extremely expensive. A much more practical strategy is to biochemically profile hundreds or thousands of small molecules (molecular weight [MW] <1000 Da) and to screen for changes in the relative levels of those compounds. By comparing two conditions, a "profile" of the differences can be obtained that is then used as a blueprint to identify the individual compounds affected.

High-performance liquid chromatography (HPLC) and gas chromatography (GC) are the most widely used analytical techniques for the separation of small metabolites ***(8)***. GC is used to separate compounds on the basis of their relative vapor pressure and affinities for the stationary phase in the chromatographic column. GC tends to give much greater chromatographic resolution than HPLC, but has the disadvantage of being limited to compounds that are volatile and heat stable. A big advantage of GC is that it can be easily combined with mass spectrometry (MS), which greatly increases its utility for multicomponent profiling because of its inherent high specificity, high sensitivity, and positive peak confirmation. Here, we describe specific protocols that were used to combine the expression of putative transcription factors in plant-cultured cells with high-throughput metabolic profiling and show the utility of this approach in investigating the function of such regulators of plant metabolism (**Fig. 1**).

2. Materials

2.1. Chemicals

2.1.1. Chemicals for GC/MS and HPLC

1. HPLC-grade methanol (Fisher Scientific).
2. HPLC-grade water (Fisher Scientific).
3. Ammonium formate (Fisher Scientific).
4. Formic acid (Sigma).

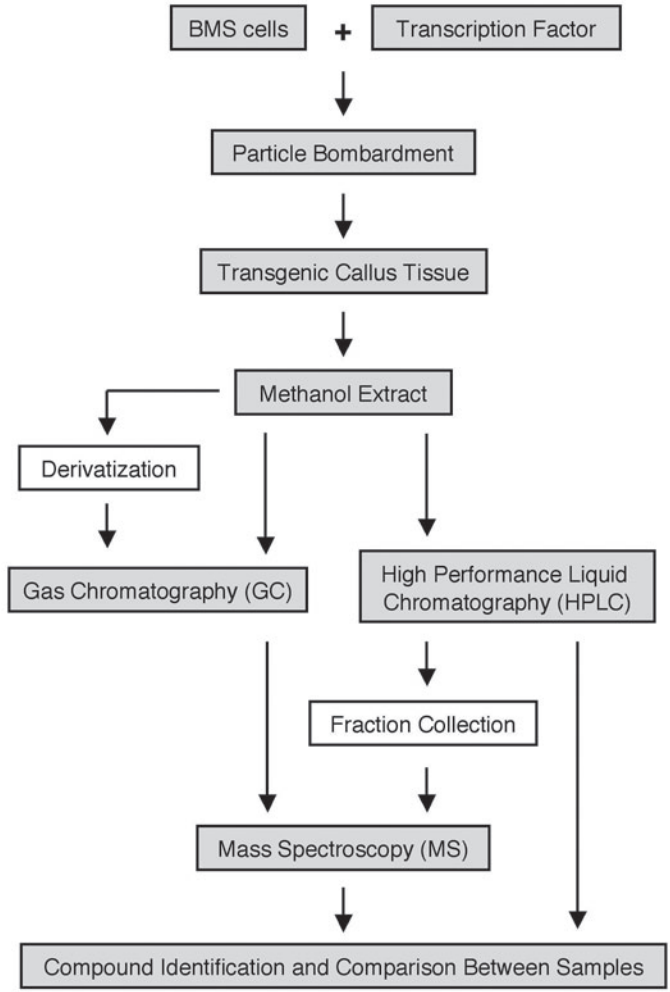


Fig. 1. Flowchart showing how metabolite profiling could be used for functional characterization of transcription factors. Different approaches that are used for dissecting biochemical networks described in the text are boxed in grey.

2.1.2. Composition of Selective Media for Maize Cell Cultures

1. Solid medium (1 L): 4.3 g Murashige and Skoog salts (Life Technologies), 0.10 g myoinositol, 30.00 g sucrose. Adjust pH to 5.6 with 1 *M* KOH, add 5.00 mL MS vitamins solution (*see* **step 2**), and 2.00 mL 2',4'-dihydro-acetophenone (0.5 g/L) (Sigma) (*see* **Note 1**). Add water to 1 L, 3.00 g phytagel (Sigma), autoclave for 20 min, cool down to about 55°–65°C, and add 3 mL phosphoinothricin (1 mg/mL) (Sigma).

2. MS vitamin solution (1 L): 0.1 g nicotinic acid, 0.02 g thiamine-HCl, 0.1 g piridoxine-HCl, 0.4 g glycine, add water to 1 L and sterilize by filtration.

2.1.3. Particle Bombardment

1. 1.5 *M* $CaCl_2$.
2. 0.1 *M* Spermidine (free base, tissue culture-grade) (Sigma).
3. 70 and 100% HPLC-grade ethanol (Fisher Scientific).
4. 3% Polyethylene glycol (8000 MW).
5. Maize expression vectors include the following parts: cauliflower mosaic virus (CaMV) 35S promoter, tobacco mosaic virus (TMV) Ω' leader, maize first *Adh1*-S intron in the 5' untranslated region (UTR), potato proteinase inhibitor (*pin* II) termination signal ***(9)***.

2.2. Equipment

2.2.1. General Equipment

1. Laminar flow cabinet.
2. Rotatory shaker.
3. Autoclave.
4. Electronic balance.
5. pH Meter.
6. Magnetic stirrer.

2.2.2. Equipment for Cell Disruption

1. 2-mL Vials (BioSpec Products; cat. no. 10832).
2. Stainless steel beads 2.3 mm (BioSpec Products; cat. no. 11079123).
3. Beadbeater™-8 (BioSpec Products).

2.2.3. Equipment for Particle Bombardment

1. Biolistic® PDS-1000/He particle delivery system (Bio-Rad).
2. Biolistic® macrocarriers (Bio-Rad).
3. Rupture disks (Bio-Rad; cat. no. 1652329).
4. Stopping screens (Bio-Rad; cat. no. 1652336).
5. 1.0-μm Gold particles (Bio-Rad).

2.2.4. Equipment for GC/MS

1. Finnigan Trace 2000 GC with split/splitless injector and Finnigan Trace MS with dedicated electron ionization (EI) ion source (Thermo Electron Corp.).
2. 30-m, 0.32 mm Internal diameter (ID), 0.25 μm XTI®-5 column (Restek).
3. 10-m Precolumn with same packing (Restek).
4. Deactivated postcolumn in the GC/MS transfer line (Restek) (*see* **Note 2**).
5. 2-mL Clear glass sample vials with open-top screw polypropylene caps and red polytetrafluoroethylene (PTFE) silicone septa (Sigma; cat. no. 27531).

6. 0.20-mL Polypropylene conical inserts with bottom spring (Sigma; cat. no. 24722).

2.2.5. Equipment for HPLC

1. Alliance® 2690 Separations Module (Waters).
2. Model 996 Photodiode Array (PDA) Detector (Waters) (*see* **Note 3**).
3. XTerra™ 4.6 × 150 mm RP-18 ODS 5 μm C18 reverse phase packing column (Waters; cat. no. 186000493) (*see* **Note 4**).
4. XTerra™ 3.9 × 20 mm guard column containing the same packing (Waters; cat. no. 186000662).
5. Universal guard holder (Waters; cat. no. WAT046910).
6. 12 × 32 mm Glass screw neck sample vials (Waters; cat. no. 186000273).
7. Screw polypropylene caps with LectraBond™ containing PTFE/silicone septa (Waters; cat. no. 186000274).
8. 0.15-mL Polypropylene conical inserts with bottom spring (Waters; cat. no. WAT094171)

3. Methods

3.1. Generation and Maintenance of Maize Callus Cell Lines

3.1.1. Preparation of Maize Callus Cells for Transformation

1. Maize Black Mexican Sweet (BMS) cells were maintained in Murashige and Skoog medium containing 2,4-D at 0.5 g/L as a suspension culture in liquid medium shaken at 150 rpm and were grown in the dark at 27°C.
2. Dry cell mass was weighed in ethanol-sterilized weighing boats. Cells were returned to a flask, and a 3-mL aliquot of liquid medium were added per gram of cells.
3. One milliliter of 50% polyethylene glycol (8000 MW) was then added to each 30-mL aliquot of medium.

3.1.2. Transformation of Maize Cells

3.1.2.1. Preparing Plates for Bombardments

1. Prepare 60 × 20 mm plates with solid media and use 3 plates per treatment.
2. Pipet 0.5 mL (approx 100 mg) of cells onto center of plate about size of a quarter. Remove extra liquid using a sterile pipet.

3.1.2.2. Coating Microcarriers

1. Forty-five microliters of microcarriers (1.0 μm) can be used for 6 bombardments, which make up one treatment. Microcarriers are stored in 55% glycerol.
2. Soak macrocarriers in 70% ethanol and dry on sterile filter paper.
3. While vortex mixing, vigorously add in order: 10 μg (10 μL) DNA, 50 μL 2.5 *M* CaCl, and 20 μL 0.1 *M* spermidine.
4. Continue vortex mixing for 2 to 3 min and allow microcarriers to settle for 1 min.

5. Spin 2 s at 14,000*g*, remove liquid, and discard.
6. Add 140 µL 70% ethanol, remove liquid, repeat with 100% ethanol, add 48 µL of 100% ethanol, mix by pipeting up and down, and place 8 µL on the center of a dry sterile macrocarrier soaked in 70% ethanol. Proceed with the bombardments immediately.

3.1.2.3. Performing Bombardments

1. Take out the rupture disk-retaining cap and soak in 70% ethanol and dry. Wipe microcarrier launch assembly, target carrier shelf and inside of chamber with 70% ethanol.
2. Soak stopping screens in 70% ethanol and dry on sterile filter paper.
3. Place a sterile stopping screen on the stopping screen support and place macrocarrier with dried DNA facing down, towards the stopping screen.
4. Wet rupture disk with 70% isopropanol for a few seconds and place in the recess of the retaining cap while still wet.
5. Install the macrocarrier holder on the top rim of the fixed nest.
6. Place the microcarrier launch assembly in the top slot and tighten to the end of the gas acceleration tube inside bombardment chamber.
7. Place the target shelf at the desired level inside the bombardment chamber. Place the Petri plate containing BMS cells on the target shelf aligning the sample to the center.
8. Close and latch the sample chamber door.
9. Turn on the vacuum source, set the vacuum switch on to the VAC position, and evacuate the sample chamber to at least 5 in of mercury. The red control switch furthest right (the Fire switch) will be illuminated when the minimum vacuum is achieved.
10. When the desired vacuum level is reached, hold the chamber vacuum at that level by quickly pressing the vacuum control switch through the middle Vent position to the bottom Hold position.
11. When the vacuum level in the bombardment chamber is stabilized, press and hold the Fire switch to allow helium pressure to build inside the gas acceleration tube that is sealed by a rupture disk.
12. Estimate rupture disk burst pressure by observing the helium pressure gauge at the top of the acceleration tube. A small pop will be heard when the rupture disk bursts, which would be within 11–13 s after the indicated rupture pressure.
13. Release the Fire switch immediately after the disk ruptures and release the vacuum in the sample chamber by setting the vacuum switch to the middle Vent position.
14. After the vacuum is released and the vacuum gauge reads 0 in of mercury, open the sample chamber door and remove the target shelf along with the sample plate.
15. Remove the microcarrier launch assembly, unscrew the lid and remove the macrocarrier holder, and discard the used macrocarrier and the stopping screen.
16. Unscrew the rupture disk-retaining cap from the gas acceleration tube and remove the remains of the rupture disk. All the apparatus is now ready for the next bombardment.

3.2. Screening for Transgenic Calli

1. The co-bombardment of the gene of interest cloned in the 35S expression vector and 35S::bialaphos-resistance gene (BAR) is carried out at a 1:1 ratio.
2. After bombardment, cells are transferred to fresh solid media, resuspended in a small vol of liquid medium after 48 h, and replated on solid media containing 3 mg/L Basta.
3. Basta-resistant calli are identified between 4 and 6 wk later and maintained on selective medium in the dark (*see* **Note 5**).

3.3. Extraction of Phytochemicals

3.3.1. Methanol Extraction of Transgenic Cell Lines

1. A sample of 100–500 mg of tissue (fresh weight) is ground to a fine suspension using 10–15 metal beads in HPLC-grade methanol (100 µL of methanol for 200 mg of tissue) in a preweighed polypropylene microfuge tube for 1 min using the Beadbeater.
2. The extracts are centrifuged (13,000*g* for 15 min) to pellet insoluble debris and the supernatant concentrated using a SpeedVac® (Savant Instruments) at a temperature below 35°C.
3. Acid hydrolysis is carried out by treating the sample with 2 *M* HCl and boiling the material in a water bath for 20 min.

3.3.2. Normalization of Methanol Extracts

1. Dry methanol extracts are weighed using an electronic balance, and the pellet is completely redissolved in HPLC-grade methanol in 10 µL/mg ratio (*see* **Note 6**).
2. Methanol extracts are centrifuged at 13,000*g* for 5 min to remove any residual debris, and the supernatant is used for GC/MS and HPLC analyses.

3.4. GC-Coupled MS

1. One microliter of sample was injected in splitless mode using an injector temperature of 250°C. The splitless time was held for 0.8 min, after which the split flow was set at 70 mL/min using a continuous septum sweep of about 5 mL/min.
2. The initial column temperature of 40°C was held for 1 min before being ramped at 11°C /min up to 310°C. Throughout the analytical portion of the run, the carrier gas flow was set at 1.0 mL/min. At the end of the analytical ramp, the flow was increased to 2.0 mL/min, and the temperature was rapidly increased to the maximum recommended for this column, 360°C, for a few minutes to condition the column for the next run.
3. A solvent delay of 5 min was used before running the MS in continuous full scan mode over the mass range of 39–459 atomic mass unit (amu) at a rate of 2.5 scans/s (*see* **Note 7**) to obtain the total ion chromatogram (TIC) (*see* **Note 8**).
4. The data analysis was carried out using the Xcalibur software (*see* **Note 9**) containing the NIST MS database and library search software (*see* **Note 10**).

3.5. HPLC

1. Solvent gradient was a modification (Blodgett and Bonello, unpublished) of the method of Rosemann et al. ***(10)***. Solvent A: 0.1% (w/v) ammonium formate in 2% (v/v) formic acid; and Solvent B: 90% (v/v) methanol and 0.1% (w/v) ammonium formate in 2% (v/v) formic acid, were used as follows: 100% solvent A isocratically for 2 min, 1 mL/min flow rate; linearly down to 90% solvent A at 4 min; 52% solvent A at 20 min; 100% solvent B at 38 min; 100% solvent B isocratically for 1 min, 1.75 mL/min flow rate; 100% solvent A at 41 min; 100% solvent A isocratically for 2.8 min; and finally 100% solvent A at 44 min, 0.5 mL/min flow rate (*see* **Note 11**).
2. The column eluate was monitored at 280 and 308 nm using the multichannel PDA detector.
3. Five microliters of sample was injected with sample temperatures maintained at 4°C and the column temperature at 30°C.
4. Identification of phenolic compounds are confirmed by co-chromatography with authentic standards (Sigma) and by retention time and spectral match using Millenium[32] PDA software (Waters) (*see* **Note 12**).

3.6. Additional Instructions for GC/MS and HPLC

1. In complex samples, the chromatographic peak for an individual component can slightly shift or be interfered with by nearly co-eluting species. Therefore, to verify the component identity, the data need to be manipulated using scan averaging and subtraction to produce a clean representative mass spectrum for comparison.
2. The relative intensity of each component can be calculated by dividing its intensity by the sum of the intensities of all of the other found compounds or of all of the other components that are not in an exclusion list. These "semiquantitative" relative intensities should be largely independent of uncontrollable variations in dilution of the extracts and should be the best way to track changes in concentrations of individual components from sample to sample without running any high precision quantification methods.
3. Unlike EI, electrospray ionization (ESI) generally produces only pseudomolecular ions (analyte molecules made ionic by association with protons or metal ions from solution) for the individual small molecules that would be observed by HPLC-MS. Since small metabolite analyses require more information and specificity than the singly charged pseudomolecular ion could provide, HPLC instrumentation and methods have been refined over time to make them more compatible with ESI-MS, due to the great popularity of tandem mass analyzers like triple quadrupole and quadrupole time of flight (TOF) hybrids. More recently, the same developments in instrumentation mentioned for GC/MS above have made accurate mass HPLC-ESI-TOF-MS instruments available. These new HPLC-MS instruments, along with the automated component detection (ACD) software should become very powerful tools for extending metabolite profiling to more polar and larger molecules.

4. Another important parameter is the accurate mass assignment. TOF mass analyzers are now commercially available, allowing mass resolution much higher than that required to measure nominal or integer mass. This high resolution is advantageous in making high precision mass assignments for the ions detected. The speed of the ion detection electronics has been the limiting factor in mass resolution and mass assignment precision until recently, due to the higher cost. However, commercial instruments are beginning to be available at a reasonable cost, which give all of the advantages of nominal mass TOF instruments, with the added benefit of accurate mass assignment down to 5 ppm. An accurate mass measurement on each ion in the GC/MS data set adds a whole new dimension to the analysis. Mass measurements this precise allow for the calculation of possible empirical formulas for molecular and fragment ions in the clean mass spectrum provided by ACD. This will greatly improve the specificity and selectivity of GC/MS analysis. Accurate mass assignment will also improve sensitivity, since narrower mass windows in selected ion chromatograms can greatly reduce the chemical noise and allow detection of smaller signals in complex mixtures.

4. Notes

1. The 2,4-D solution tends to crystallize upon storage at 4°C. To prevent this, the solution can be prepared in 50% ethanol in water.
2. Using smaller ID and shorter columns, higher carrier flow rates and faster temperature ramping programs can greatly shorten the time required to produce the GC/MS data sets required for metabolite profiling. Useful fast chromatography methods for this application would still have fairly high chromatographic resolution. Therefore, not only are the runs shorter (typically by more than a factor of ten), but the time it takes for an individual component peak to elute (the peak width) also becomes a lot narrower (by as much as a factor of ten). As a result, the mass spectrometer must acquire individual mass spectra at a faster rate to continue to define the peak shape.
3. The PDA detector has a wavelength range from 190–800 nm with an accuracy of ±1 nm.
4. XTerra columns containing the bonded and end-capped 5-μm particles gives the highest most homogeneous coverage, allowing best peak shape and a pH range from 1–12.
5. The transgenic cell lines were subcultured onto new medium containing 0.5 g/L 2,4-D every 4 wk.
6. If the methanol pellet becomes too dry, resuspension can be facilitated by using 50% HPLC-grade methanol.
7. The quadrupole mass spectrometer was set at 2.5 scans/s from mass 39–459, to make the 3-s wide peaks about 7.5 scans wide at half height. One disadvantage of a quadrupole or sector mass spectrometer is the fact that it only transmits one resolution element or mass at a time through their mass analyzers. To produce a mass spectrum, the analyzer is scanned through the selected range, and the ion intensities are measured sequentially. The total ion current in the mass spectrum

of an individual component changes as the molecular flux for that component changes in the ion source of the mass spectrometer. The GC peak for an individual component is a nearly continuously changing flux when the component is eluting. So, the slow scans of a sequentially scanned mass analyzer will produce different mass spectra depending on whether the scan occurred during the rising, apex, or falling part of the signal. This characteristic makes automatic peak detection harder to accomplish as the peaks become narrow compared to the scan speed available.

8. The reconstructed TIC represent the volatile portion of the metabolite profile from these extracts. Gross changes in the profile can be observed by simply comparing these TICs using the same time base for the data collected from samples representing different conditions of gene expression. The accumulation or depletion of metabolite species will show up as changes in the intensities of corresponding chromatographic peaks in the TIC. The advantage of using GC/MS is that there is a full-scan mass spectrum associated with each point in the TIC. The presence and relative intensities of the ions produced as each component elutes from the GC column during the run are recorded in these mass spectra. These mass spectra are representative of the components eluting from the column at each point in the run and can be used to verify that peaks in similar locations in different runs represent the same metabolite molecule. Conversely, a component found in one sample can be searched for in the data from a second sample by searching the corresponding region of the chromatogram for mass spectra with the same ion intensity patterns.
9. The Xcalibur data analysis software produces reconstructed ion chromatograms using any user-selected combination of characteristic ions. There is a vast amount of information in the GC/MS data. Selected ion chromatograms can often represent homologous series of related species in the data, such as free fatty acids, α-substituted fatty acid methyl esters, alkyl-substituted phenols, and many other such specific profiles.
10. The EI mass spectra of individual metabolite species tend to be characteristic and are often quite unique. This has long been recognized and is taken advantage of by the publication of libraries or databases of high quality EI mass spectra associated with specific molecular species. These metabolite profiles, like many natural extracts, tend to be rather complex. Below the level of the major components, which may be observed as distinct peaks in the chromatogram, there is often a nearly continuous series of minor components that are present at too low a level to show up as a peak in the TIC. GC/MS data analysis packages, like Xcalibur, allow the operator to interactively select scans or average ranges of scans and subtract scans or multiple ranges of nearby scans to produce "clean" mass spectra for library search identifications or for comparison to components found in runs from other samples. A major drawback is that this is very labor-intensive and makes it hard to take full advantage of the specificity and sensitivity of the GC/MS data.
11. The Separation Module is degassed, the solvent management system primed, and the sample management system purged every time after turning the HPLC equipment on and before running samples.

12. Compound identification in HPLC is carried out by comparing the retention times and spectra to standards that are run in the same chromatographic conditions. The limitations in this includes the number of standards that are available for making a user library, which would take a prohibitive amount of effort in metabolite profiling, since we are looking at the difference between chromatographic profile of a control and a test sample. One way of identifying peaks that do not correspond to any standards in a user library is by separating the peaks using a fraction collector and then subjecting them to a direct electrospray MS analysis. The specificity and speed of HPLC analysis can be increased by coupling it to a mass spectrometer. ESI can be used to obtain mass spectra from the components that elute from the HPLC column.

References

1. Braun, E. L., Dias, A. P., Matulnik, T. J., and Grotewold, E. (2001) Transcription factors and metabolic engineering: novel applications for ancient tools. *Recent. Adv. Phytochem*. **35,** 79–109.
2. Memelink, J., Kijne, J. W., van der Heijden, R., and Verpoorte, R. (2001) Genetic modification of plant secondary metabolite pathways using transcriptional regulators. *Adv. Biochem. Eng. Biotechnol.* **72,** 103–125.
3. Grotewold, E., Chamberlain, M., St. Claire, G., et al. (1998) Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. *Plant Cell* **10,** 721–740.
4. Dias, A. P. and Grotewold, E. (2003) Manipulating the accumulation of phenolics in maize cultured cells using transcription factors. *Biochem. Eng. J.*, **14,** 207–216.
5. Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol*. **18,** 1157–1161.
6. Roessner, U., Luedemann, A., Brust, D., et al. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13,** 11–29.
7. Cannell, R. J. P. (ed.) (1998) *Natural Products Isolation. Methods in Biotechnology*. Humana Press, Totowa.
8. Glassbrook, N. and Ryals, J. (2001) A systematic approach to biochemical profiling. *Curr. Opin. Plant Biol*. **4,** 186–190.
9. Grotewold, E., Drummond, B., Bowen, B., and Peterson, T. (1994) The Myb-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell* **76,** 543–553.
10. Rosemann, D., Heller, W., and Sandermann, H., Jr. (1991) Biochemical plant responses to ozone II. Induction of stilbene biosysthesis in scots pine (*Pinus sylvestris* L.) seedlings. *Plant Physiol.* **97,** 1280–1286.

26

Growth Stage-Based Phenotypic Profiling of Plants

Susanne Kjemtrup, Douglas C. Boyes, Cory Christensen, Amy J. McCaskill, Michelle Hylton, and Keith Davis

Summary

Recent high-throughput methods for the analysis of biological samples are raising the possibility of data integration between investigators and even across technology platforms. In plant biology, integration of data can be problematic, since heterogeneity of plant growth conditions and phenotypes result in data sets that are not consistent or easily comparable. In this chapter, we describe the development of a plant phenotyping platform based on a growth stage scale that will aid in the generation of coherent data. While the emphasis is on the development of a phenotyping platform for *Arabidopsis*, the aim of this chapter is to describe principles that can be applied to other plant systems as well. Additionally, we discuss approaches for data analysis and quality control.

Key Words

high-throughput, *Arabidopsis*, phenotyping platform, phenotype, growth stage, BBCH, phenotype taxonomy, phenomics, LIMS

1. Introduction

Recent genomic technology initiatives, such as the *Arabidopsis* 2010 project ***(1)***, have the goal of defining the function of plant genes on a large scale. A requisite part of gene function determination is the characterization of phenotypes that result when the expression level of the gene of interest is altered. The resultant phenotypes may be molecular, biochemical, morphological, or developmental in nature and are often associated with one or more stages of growth.

Plants exhibit widely different developmental timelines and morphologies depending on the environment in which they are grown. Thus, comparison of data collected by laboratories in which plants are grown under slightly differ-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

ent conditions can be problematic. This is especially true if the data are collected solely with reference to chronological age. In contrast, phenotypic data collected with reference to a commonly defined series of growth stages would provide a coherency not otherwise achieved by collection procedures based on chronological age alone.

Common growth stage definitions have been developed for a number of experimental organisms, including *Caenorhabditis elegans* and *Drosophila* ***(2,3)***. Similar scales have also been developed for many agronomically important plant species (e.g., a decimal code for the growth stages of cereals ***[4]***). One of these is the BBCH growth stage scale, named for the consortium of agricultural companies that developed it (Basf, Bayer, Ciba-Geigy, and Hoechst). The BBCH scale provides a comprehensive growth stage description and can be adapted for most crop and weed species ***(5)***. Originally developed as a means of communication among agriculturists, adaptation of such a universal scale at the laboratory level would allow for easier data comparisons between and within species. The use of growth stage definitions will greatly facilitate information sharing and increase the value of individual research projects.

The BBCH scale assigns a numerical value (0–9) to 10 principal developmental stages that occur throughout plant development: 0, germination, sprouting; 1, leaf development; 2, formation of side shoots; 3, stem elongation–rosette growth; 4, vegetative plant parts; 5, inflorescence emergence; 6, flowering; 7, fruit development; 8, ripening; 9, senescence. Each principal growth stage is subdivided into 10 more detailed morphological events germane to the principal stage. The resulting code provides a digital naming convention for nearly any developmental stage of a plant at any given time (*see* **Table 1**).

We have adapted a modified version of the BBCH scale for high-throughput phenotyping of *Arabidopsis* ***(6)***. This chapter describes a two-phase method for the collection of data for both quantitative and qualitative traits spread over the developmental timeline of the plant. In the first phase of the method, data is collected, enabling a series of landmark growth stages to be defined. The second phase involves the collection of detailed data for additional traits that are of particular interest at any one of these given stages. **Figure 1** illustrates the growth stages and phases we use for data collection. While we focus on the application of this method to *Arabidopsis*, a similar strategy can be applied to the collection of similar data from other plant species as well.

Table 1
BBCH Growth Scale *(5)*

Numeric code	Growth stage description
0	**Germination; sprouting.**
00	Dry seed.
01	Seed imbibition begins.
03	Seed imbibition complete.
05	Radicle emerged from seed.
06	Elongation of radicle, formation of root hairs or lateral roots.
07	Coleoptile emerged; hypocotyls with cotyledons broken through seed coat.
08	Hyocotyl with cotyledons grow toward soil surface.
09	Cotyledons or coleoptile breaks through soil surface.
1	**Leaf development (main shoot).**
10	First true leaf emerged from coleoptile; or cotyledons completely unfolded.
11	First true leaf.
12	Two true leaves.
13	Three true leaves, etc.
19	Nine or more true leaves (if tillering or shoot and stem elongation occur at an earlier stage or not at all, continue with either stage 21 or 31).
2	**Formation of side shoots or tillering.**
21	First side shoot or tiller visible.
22	Two or more side shoots or tillers visible.
23	Three or more side shoots or tillers visible, etc., to 28.
29	Nine or more side shoots or tillers visible.
3	**Stem elongation or rosette growth (main shoot; shoot development).**
31	Stem (rosette) 10% of final length (diameter) or 1 nodes detectable.
32	Stem (rosette) 20% of final length (diameter) or 2 nodes detectable.
33	Stem (rosette) 30% of final length (diameter) or 3 nodes detectable, etc., to 38.
39	Maximum stem length or rosette diameter reached; 9 or more nodes visible.
4	**Development of harvestable vegetative plant parts.**
41	Harvestable vegetative plant parts begin to develop or flag leaf sheath extending.
43	Harvestable vegetative plant parts have reached 30% of final size; or flag leaf sheath just visibly swollen.
45	Harvestable vegetative plant parts have reached 50% of final size; or flag leaf sheath swollen.

Table 1
Continued

Numeric code	Growth stage description
47	Harvestable vegetative plant parts reach 70% of final size; or flag leaf sheath opening.
49	Harvestable vegetative plant parts reach final size; or first awns visible.
5	**Inflorescence emergence (main shoot); ear or panicle emergence.**
51	Inflorescence or flower buds visible.
55	First individual (closed) flowers visible.
59	First flower petals visible; or inflorescence fully emerged.
6	**Flowering on main shoot.**
61	Beginning of flowering: 10% flowers open.
63	30% Flowers open.
65	50% Flowers open; first petals fallen or dry.
67	Flowering finishing; majority of petals fallen or dry.
69	End of flowering; fruit set visible.
7	**Development of fruit.**
71	Small fruits visible or fruit has reached 10% of final size.
73	First fruits have reached final size or fruit has reached 30% of final size.
75	50% Fruits have reached final size or fruit has reached 50% of normal size.
77	70% Fruits have reached final size or fruit has reached 70% of normal size.
79	Nearly all fruits have reached final size.
8	**Ripening or maturity of fruit and seed.**
81	Beginning of ripening or fruit coloration.
85	Advanced ripening or fruit coloration.
87	Fruit begins to soften.
89	Fully ripe; beginning of fruit abcission.
9	**Senescence: beginning of dormancy.**
91	Shoot development completed; foliage still green.
93	Leaves begin to change color or fall.
95	50% Leaves discolored or fallen.
97	Plant material dead or dormant.
99	Harvested seed.

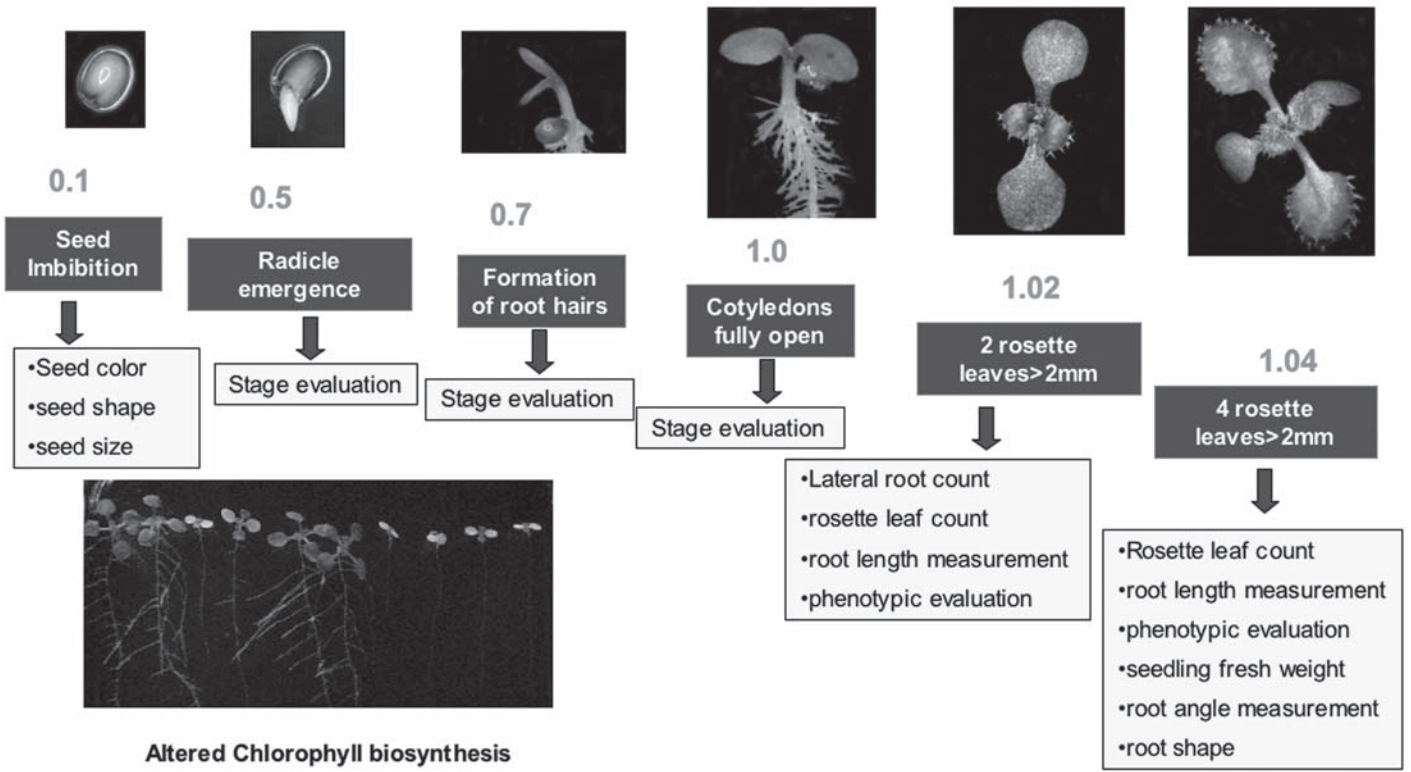


Fig. 1. **(A)** Growth stage-based early analysis of *Arabidopsis*: growth stages and descriptions are indicated in the black box. Additional data collection steps are outlined in the boxes below each stage.

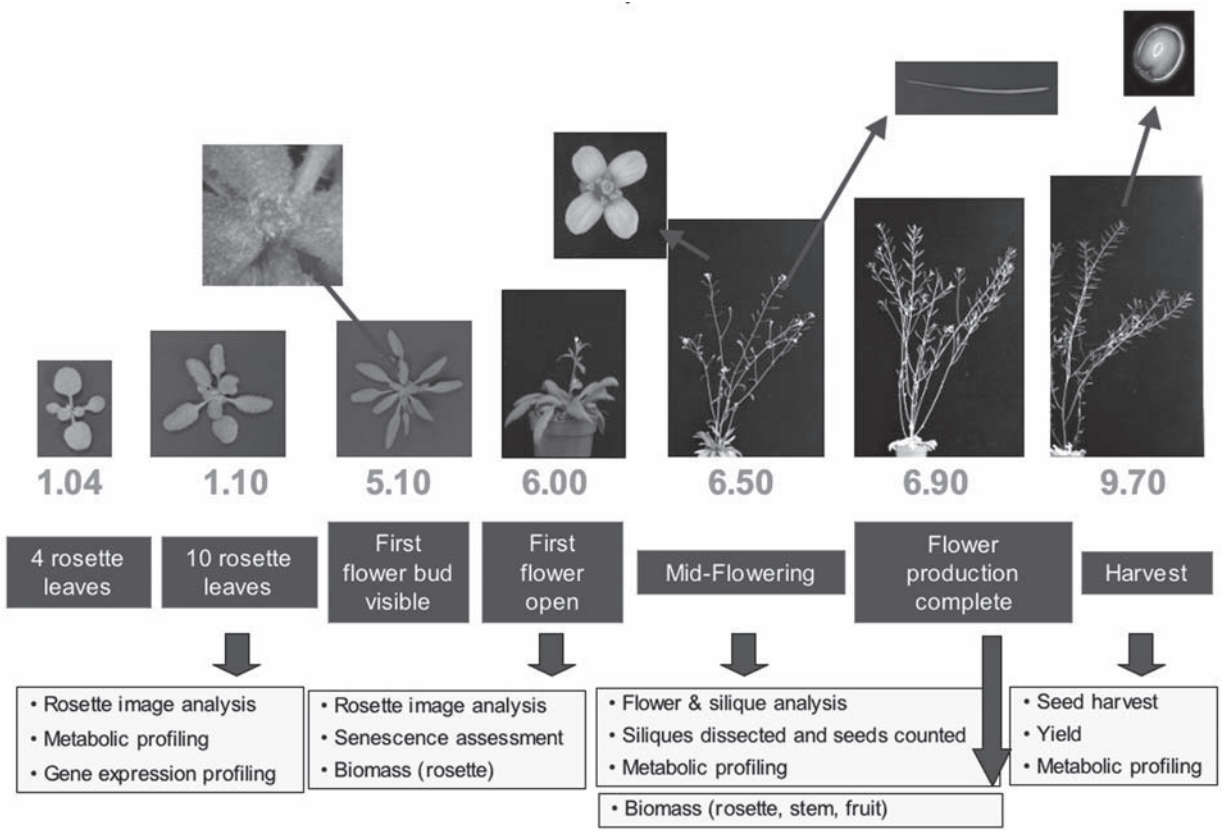


Fig. 1. **(B)** Growth stage-based analysis of soil-grown *Arabidopsis*: growth stages and descriptions are indicated in the black box. Additional data collection steps are outlined in the boxes below each stage.

2. Materials

2.1. Arabidopsis *Growth and Maintenance*

1. Growth chambers: Model TCR 480 (Conviron). Each of these chambers is outfitted with 20 growth racks. Each rack has four shelves of growth space, each of which is approx 2 × 4 ft in size and accommodates four standard greenhouse flats. Thus, the maximum capacity of an entire chamber is 320 flats.
2. Potting medium: e.g., Metro Mix 360 (Scotts).
3. Granular time-release fertilizer: e.g., Osmocote 18–6–12 (Scotts).
4. Mini-Mayer 2100 potting machine (Gro-May) modified to fill 2 in square pots (*see* **Note 1**).
5. Pipets or liquid handling robot (e.g., Genesis RSP-200; Tecan US) to sow seed (*see* **Note 2**).
6. 0.1% (w/v) Agarose solution in water to suspend seeds for sowing.
7. Cold room or refrigerator to stratify seeds at 4°C.

2.2. Data Collection Instruments

1. 300-mm Ruler (VWR Scientific).
2. Electronic or manual counters (VWR Scientific).
3. Electronic calipers (Mitutoyo America) connected to the computer via an RS232 port. Alternatively, manual calipers (VWR Scientific) will also suffice.
4. Digital camera with USB or firewire connection to computer. We use Nikon® D1 series cameras (Nikon).
5. Balance or automated weighing station.

2.3. Data Collection Software

Options for data collection software include:

1. Electronic spreadsheets (e.g., Microsoft® Excel®).
2. Customized relational databases (e.g., Microsoft Access, ORACLE).
3. Commercial electronic laboratory notebooks (ELN). Some popular ELNs include LabTrack (http://www.labtrack.com/), which is by far the most popular ELN available. It acts as a laboratory information management system (LIMS) and a notebook. This ELN acts not only as a word processor, but also provides the reporting and searching capabilities of a relational database. An ELN can also become a legally acceptable document with the addition of service subscriptions like First Use (http://www.firstuse.com/) or Surety (http://www.surety.com/).

2.4. Data Analysis Software

2.4.1. Data Analysis: Commercial

Commercial options for data analysis software include:

1. Statistical analysis software by SAS®.
2. Microsoft Excel.

2.4.2. Image Analysis: Commercial

Commercial options for image analysis software include:

1. Image Pro Plus (http://www.mediacy.com/ippage.htm) and (http://www.optimas.com/optimas.htm) offered by Media Cybernetics.
2. IP Lab (http://www.scanalytics.com/product.html) for Macintosh® and Windows™ operating systems offered by Scanalytics.

2.4.3. Image Analysis: Public Domain

There are also a number of public domain software packages available for image analysis, including:

1. National Institutes of Health (NIH) Image for Macintosh® (http://rsb.info.nih.gov/nih-image/) and ImageJ for any computer with Java 1.1 (http://rsb.info.nih.gov/ij/).
2. Scion Image (http://www.scioncorp.com/frames/fr_scion_products.htm), which is a windows equivalent to NIH Image.
3. Image Tool (http://ddsdx.uthscsa.edu/dig/itdesc.html).

3. Methods

3.1. Arabidopsis *Growth and Maintenance*

3.1.1. Growth Conditions

Large-scale phenotyping efforts require consistent long-term reproducibility of environmental conditions for plant growth. Unlike greenhouses, growth chambers are subject to fewer outside environmental influences and are easier to control. Growth chambers have the added benefit of allowing plants to be grown at a higher density than is possible in most greenhouses. We maintain the following conditions in our chambers:

1. Lighting: the light intensity over each shelf is maintained at 175–200 μE by a fixture containing nine T8 cool white fluorescent tubes. The distance between the fluorescent tubes and the shelf below is approx 55 cm. To ensure consistent illumination over time, one-third of the fluorescent tubes are replaced every 3 mo.
2. The day length is 16 h.
3. Daytime temperature is 22°C.
4. Nighttime temperature is 20°C.
5. Relative humidity is held constant at 65%.

3.1.2. Potting Medium Preparation

1. One bag (3 cu ft) of commercial potting medium is supplemented with 90 g of granular fertilizer and 2 gallons water.

2. Components are mixed until evenly dispersed using a cement mixer or commercial soil mixer.
3. Pots are filled manually, or for high-throughput operation, filled with a potting machine (*see* **Note 1**).
4. A standard 10 in. × 20 in. greenhouse flat holds 32 2-in pots configured in a 4 × 8 grid.

3.1.3. Seed Sowing

1. Prior to sowing, seeds are suspended in a solution of 0.1% (w/v) agarose and placed at 4°C for 3 d to synchronize germination.
2. Depending on throughput and application, sowing can be performed either manually using a pipet, or in a more automated fashion, using a liquid handling robot (*see* **Note 2**).
3. After sowing, the flats are watered, covered with a humidity dome, and placed in the growth room.
4. Following germination, the humidity dome is removed, and the flats are irrigated every 2 d until mid-flowering, at which point watering is increased to every day, until seed set is complete.
5. Flats are irrigated from below using an ebb and flood method. All irrigation water is purified by reverse osmosis prior to use.

3.2. Collection of Data to Define Growth Stages

1. Growth-stage scale development or determination: the BBCH scale is a ready-made template to aid in the definition of landmark growth stages. However, its generic nature requires that it be more inclusive than exclusive, making the challenge of adapting it to a particular plant species one of detail reduction and focus. The scale we use for *Arabidopsis* is based on a version of the BBCH scale already developed for the related plant, *Brassica* ***(5)***. Principal growth stages 2 and 4 were removed from this version of the scale, as they reference tiller formation and harvestable seed production in monocots. The remaining principal growth stages of relevance are 0, 1, 3, 5, 6, 7, 8, and 9.
2. Scale refinement (if necessary): further refinement of growth stage definitions may be required for operational use, especially for those cases where growth stages are defined relative to the percentage of completion of that stage. An example of this is principal growth stage 6 (inflorescence development). The beginning (stage 6.0) and end of inflorescence development (stage 6.9) are easily identified in real time, as the time to first flower opening and flowering completion, respectively. In contrast, the point at which 50% of the inflorescence has been produced (stage 6.5) can only be defined in retrospect, after flowering is complete. Thus, the implementation of growth stages that are defined relatively as real-time data collection triggers is a practical impossibility. A useful strategy in these cases is to identify a trait that can be followed as a surrogate to determine when the growth stage of interest has been reached. For example, in a pilot experiment we counted flowers every other day between stages 6.0 and 6.9, and simultaneously measured stem height. We found that the rate of stem elongation

Table 2
***Arabidopsis* Growth Stages and Measurements**

Stage	Description	Measurement or action
Growth stage 0	**Seed germination**	
0.10	Seed imbibition.	Visual inspection.
0.50	Radicle emergence.	Visual inspection.
0.7	Hypocotyl and cotyledon emergence.	Visual inspection.
R6	More than 50% of the seedlings have primary roots ≥6 cm in length.	Caliper measurement of root.
Growth stage 1	**Leaf development**	
1.0	Cotyledons fully opened.	Visual inspection.
1.02	2 rosette leaves >1 mm in length.	Leaf count.
1.03	3 rosette leaves >1 mm in length, etc., to stage 1.14.	Leaf count.
Growth stage 3	**Rosette growth**	
3.20	Rosette is 20% of final size.	Caliper measurement of longest leaf.
3.50	Rosette is 50% of final size.	
3.70	Rosette is 70% of final size.	
3.90	Rosette growth complete.	Tissue harvest.
Growth stage 5	**Inflorescence emergence**	
5.10	First flower buds visible.	Visual inspection.
Growth stage 6	**Flower production**	
6.00	First flower open.	Visual inspection; tissue image.
6.10	10% Flowers to be produced have opened.	Ruler measurement of stem height for correlative determination.
6.30	30% Flowers to be produced have opened.	
6.50	50% Flowers to be produced have opened.	Image.
6.90	Flowering complete.	Tissue dissection and dry weight.
Growth stage 7	**Silique filling**	
Growth stage 8	**Silique ripening**	
8.00	First silique shattered.	Visual inspection.
Growth stage 9	**Senescence**	
9.70	Senescence complete; ready for seed harvest.	Seed harvest.

plateaued concomitantly with stage 6.5 as defined by the flower count and could, therefore, serve as a surrogate trait. From this result, we developed a working definition of stage 6.5, as the day on which the rate of stem elongation decreased by more than 20% for two consecutive measurement cycles ***(6)***.

The growth stages we use routinely in the analysis of *Arabidopsis* phenotypes are listed in **Table 2**. Note that each growth stage is defined by a simple observation or robust measurement that can be determined rapidly (*see* **Note 3**). It is important that measurement of growth stage-determining traits be rapid and simple, because additional data specific to a particular growth stage can thus be collected concurrent with the attainment of that stage (*see* **Note 4**).

3. Characterization of traits to measure at specific growth stages: besides growth stage measurements, additional traits can be assembled into modules to collect more extensive data for a particular stage of interest. Modules could include characterization of floral morphology at mid-flowering (stage 6.5), yield and seed-related traits at the conclusion of seed maturation (stage 9.7), or disease characterization during vegetative development (stage 1.10). The scope of the data collection in these modules can range from a broad survey of traits, in an attempt to uncover as many phenotypes as possible, to the analysis of a specific trait at a single stage of growth (*see* **Note 5**). **Figure 1A,B** illustrate a variety of *Arabidopsis* traits that can be collected for early analysis on plates and whole plant analysis on soil.
4. Evaluation of possible quantitative traits: traits can be quantitative or qualitative and can include processes such as harvesting tissue samples for subsequent extraction and analysis by methods including gene expression profiling and biochemical profiling. Examples of robust quantitative traits for the analysis of *Arabidopsis* include biomass of leaves, stems, siliques, and seeds, as well as the length of siliques, pedicels, etc. These traits can be assessed easily through the use of standard equipment including a balance, caliper, and ruler (*see* **Note 6**). With a greater investment in technology, a large number of metrics can also be extracted from digital images. Traits such as area, perimeter, major and minor axis, and shape (e.g., eccentricity, standard deviation of the radius) can be quantitated readily from an image of seeds, siliques, pollen grains, or an intact rosette. Image analysis can also be used to more precisely assess traits, such as flower size, that are challenging to measure by hand (*see* **Note 7**).

   Technology can also be applied to the analysis of other traits. For instance, abnormal leaf color can be indicative of any number of underlying metabolic or developmental defects. While color can be assessed qualitatively (*see* **step 5**), one can also use a color spectrophotometer to quantify the wavelengths of light reflected from the subject.
5. Evaluation of possible qualitative traits: clearly, all visual phenotypes will not be represented equally through an assessment of the quantitative traits such as those described in **step 4**. Therefore, qualitative descriptors and images should also be included as part of the phenotyping process. While free text can be used as a means to capture descriptive data, we have developed an *Arabidopsis* Pheno-

type Taxonomy (APT) specifically for this purpose (*see* **Note 8**). The APT consists of a structural hierarchy (e.g., inflorescence::stem::flower::petal) that has modifying terms for every structure (e.g., inflorescence::height, stem::width, flower::male sterile, petal::color). The APT contains descriptions of shape, size, color, dimension, etc., for each major feature of *Arabidopsis*, as well as descriptions for altered developmental timing and stress tolerance. Pleiotropic phenotypes can be described through the assignment of multiple APT entries. For maximum utility, the APT entries can be associated with a corresponding image of the plant.

6. Efficient measurement design for a population of plants: phenotypic profiling invariably involves the analysis of populations of plants. In some cases, the population size can be very large, and even with a computerized data collection system, it quickly becomes a logistical difficulty to track the development of each plant individually and to collect growth stage-specific data for each at the appropriate time. To address this problem, we exploited the fact that the time required for individual plants to reach a growth stage is distributed normally within the population. Data to determine growth stage is collected at the level of individual plants, and the population is considered to have reached a growth stage when 50% or more of the surveyed individuals have reached the growth stage of interest. This event triggers the collection of additional data specific for that growth stage from all of the individuals within the population. This method reduces the complexity of the data collection process by providing a mechanism to schedule growth stage-specific data only once during the development of each population.

   Some of the data collection processes result in destruction of the specimen. This should be taken into account when designing the phenotyping process. When a population reaches a growth stage of interest, a subset of plants can be harvested for analysis, while the remaining plants are allowed to continue to grow for later analysis. This strategy permits a complete set of developmental data to be collected from plants grown at the same time under the same conditions.

3.3. Sample Tracking and Data Entry

Various types of software can be used to track samples and record data.

1. For small-scale experiments, a spreadsheet may suffice.
2. However, for larger studies, a relational database is preferable. Relational databases allow information to be stored and retrieved more efficiently than spreadsheets. Relational databases support the development of graphical user interfaces that enable highly efficient entry of data as well as more sophisticated queries of the data (*see* **Note 9**).

3.4. Quality Control

Variability is inherent in the assessment of biological phenomena. While phenotypic variation resulting from a genetic difference is typically a desired outcome, experiments can be compromised as result of uncontrolled variabil-

ity from undesirable sources. Some of these sources include variation in environmental conditions, data collection technique, and data entry errors. Safeguards to minimize these and other undesirable sources of variability are essential.

One or all of the following steps can be implemented to help control the quality of the data collection process:

1. Develop and adhere to standard operating procedures pertaining to those components described above to minimize phenotypic variation resulting from inconsistency in environmental conditions. For example, we routinely monitor the electroconductivity and pH of the irrigation water and utilize a plant growth bioassay to screen each lot of soil prior to use.
2. The intensive nature of high-throughput phenotyping often requires the efforts of many people in the data collection process. A thorough training program can help to reduce variation resulting from differences in the way different people collect the same types of data. Training program effectiveness can be monitored through the comparison of duplicate data sets collected by different individuals. The variation within the difference of the duplicate data sets can be taken as a representation of the variability resulting from the measurement system.
3. High-throughput environments are prone to data entry errors. These can be minimized through the incorporation of high and low data limits at the level of the data collection interface. Additional measures can be added to ensure that values entered for a trait are consistent. For example, in recording the number of rosette leaves over time, an entry with a value of 6 after a previous entry of 7 would trigger an error message.

3.5. Data Analysis

Phenotypic data resulting from growth stage-based data collection is a mixture of both quantitative and qualitative measurements. The following generic data analysis method encompasses both types of data. Be aware that, when collecting and analyzing data from a phenotypic platform, the sample size for a particular trait will vary depending on the subpopulation of plants that are sampled. There is a necessary balance between the analysis of plants at a high-throughput and collecting data with sufficient replication to detect subtle phenotypic differences with high confidence. The requirements on both sides of this equation will vary with the application and must be evaluated prior to initiating a phenotyping platform.

1. A subpopulation of control plants is included within each flat of plants grown for phenotypic analysis (*see* **Note 10**).
2. Quantitative data are averaged within the control and mutant plant populations. The mutant and control means are compared to each other using a *t*-test (*see* **Note 11**).

3. Qualitative data are represented as a frequency of the noncontrol responses. For example, a mutant mean of 0.75 for the "seedling color" component would mean that 75% of the seedlings exhibited a color that differed from the control. The frequencies of the control and mutant populations are compared using a *t*-test to determine whether they differ significantly (*see* **Note 11**).

4. Notes

1. While pots can be filled by hand, we have found that the use of a potting machine improves process efficiency and delivers more consistent soil compaction.
2. A liquid handling robot has the advantage of increasing sowing efficiency and also enables the location of controls and/or seed lines to be easily randomized within the flat.
3. During the collection of developmental data over time, it is most efficient to schedule observations with reference to data collected during the previous cycle of analysis. If the data collection is being driven by computerized system, this can be accomplished through the definition of a series of rules that trigger data collection events. For example, the transition to flowering is first observed as the production of floral buds. This event usually occurs several days prior to the opening of the first flower. If one is interested in capturing the timing of both events, the occurrence of the floral buds can serve as the trigger to begin assessing flower opening. To continue the example, opening of the first flower is correlated with the completion of vegetative development. Therefore, opening of the first flower makes an ideal trigger to cease all observations related to vegetative development (e.g., number of leaves, rosette size, etc.).
4. Planning a growth stage-based experiment is not as straightforward as planning for a calendar-based experiment; achieving a specific growth stage happens within a time frame, not necessarily on a specific day. While some of those time frames can be fairly tight (in the Col-0 ecotype of *Arabidopsis*, days to stage 1.02 has a standard deviation of 1.3 d ***[6]***), others can be fairly broad (days to stage 6.50 in Col-0 has a standard deviation of 4.9 d ***[6]***).
5. Certain measurements also require additional considerations. For example, the distance across an open flower can only be measured in the morning, because under our conditions, the flowers close in the afternoon.
6. Data collection can be adapted to a high-throughput environment through the use of balances and calipers that are connected directly to a computer for automated data entry. Even higher throughput weighing of seeds or other samples can be achieved using an automated balance workstation (e.g., Mettler-Toledo Bohdan Automation; http://www.bohdan.com/index.htm).
7. The success of digital image analysis relies on the ability to take high quality images with consistent magnification and lighting. Therefore, it is most efficient to develop one or more workstations that are dedicated to image capture. We use Nikon D-series cameras. They are available with a choice of resolution and are built on a standard 35-mm camera body, enabling an extensive array of macro and micro lens configurations.

8. The complete APT can be found on the Paradigm Genetics Web site (www.paradgimgenetics.com).
9. Relational databases can range in complexity from simple systems built in Microsoft Access or the open source database MySQL, to advanced systems built on a platform such as ORACLE. Additionally, there is a selection of ELNs available that would also suffice. We use an ORACLE-based electronic LIMS that tracks the location and status of all samples through unique 10-digit identifiers. This system also provides an advanced data collection interface, which incorporates a rule set to evaluate data immediately upon entry, to automatically determine when a growth stage has been achieved. This event then triggers the collection of additional data that are specific to that growth stage.
10. Data obtained from the control plants are critical not only as a reference for assessing phenotypic variation, but also as a means to control and understand the consistency of growth conditions within and between locations in the growth rooms. A control data set can be developed from a large group of control plants that represent a random sample of sow dates and/or growth locations. This reference control can then be used to perform *t*-tests with the control data from any individual flat in a manner analogous to that used to compare mutants to the control. The reference control may be used to help refine the analysis of mutant data as well. For example, if the control data within an experimental flat differ significantly from the reference control population, then the reliability of the data from the mutant plants in that flat should also be called into question.
11. The *t*-test value may be interpreted as the number of standard errors between the mutant mean and the control mean. The value of the *t*-test statistic may be positive or negative, representing mutant variation that is greater or less than the control mean, respectively. For sample sizes greater than three, one can be at least 95% confident that *t*-test values greater than 2 standard errors from the control are due to biological variation and not simply the result of chance.

References

1. *Arabidopsis* 2010 Project. (http://www.nsf.gov/pubs/2001/nsf01162/nsf01162.html).
2. Hartenstein, V. (1993) *Atlas of Drosophila Morphology and Development*. CSH Laboratory Press, Cold Spring Harbor, NY.
3. Wilkins, A. (1993) *Genetic Analysis of Animal Development*. Wiley-Liss, New York.
4. Zadoks, J., Chang, T., and Konzak, C. (1974) A decimal code for the growth stages of cereals. *Weed Res.* **14,** 415–421.
5. Lancashire, P., Bleiholder, H., vd Boom, T. P. L., Stauss, R., Weber, E. and Witzenberger, A. (1991) A uniform decimal code for growth stages of crops and weeds. *Ann. Appl. Biol.* **119,** 561–601.
6. Boyes, D. C., Zayed, A. M., Ascenzi, R., et al. (2001) Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* **13,** 1499–1510.