METHODS IN MOLECULAR BIOLOGY™ 387

Serial Analysis of Gene Expression (SAGE)

Methods and Protocols

Edited by
Kåre Lehmann Nielsen

Humana Press

Serial Analysis of Gene Expression (SAGE)

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

387. Serial Analysis of Gene Expression (SAGE): *Methods and Protocols,* edited by *Kåre Lehmann Nielsen, 2008*

386. Peptide Characterization and Application Protocols, edited by *Gregg B. Fields, 2007*

385. Microchip-Based Assay Systems: *Methods and Applications,* edited by *Pierre N. Floriano, 2007*

384. Capillary Electrophoresis: *Methods and Protocols,* edited by *Philippe Schmitt-Kopplin, 2007*

383. Cancer Genomics and Proteomics: *Methods and Protocols,* edited by *Paul B. Fisher, 2007*

382. Microarrays, Second Edition: *Volume 2, Applications and Data Analysis,* edited by *Jang B. Rampal, 2007*

381. Microarrays, Second Edition: *Volume 1, Synthesis Methods,* edited by *Jang B. Rampal, 2007*

380. Immunological Tolerance: *Methods and Protocols,* edited by *Paul J. Fairchild, 2007*

379. Glycovirology Protocols, edited by *Richard J. Sugrue, 2007*

378. Monoclonal Antibodies: *Methods and Protocols,* edited by *Maher Albitar, 2007*

377. Microarray Data Analysis: *Methods and Applications,* edited by *Michael J. Korenberg, 2007*

376. Linkage Disequilibrium and Association Mapping: *Analysis and Application,* edited by *Andrew R. Collins, 2007*

375. In Vitro Transcription and Translation Protocols: *Second Edition,* edited by *Guido Grandi, 2007*

374. Biological Applications of Quantum Dots, edited by *Marcel Bruchez and Charles Z. Hotz, 2007*

373. Pyrosequencing®Protocols, edited by *Sharon Marsh, 2007*

372. Mitochondrial Genomics and Proteomics Protocols, edited by *Dario Leister and Johannes Herrmann, 2007*

371. Biological Aging: *Methods and Protocols,* edited by *Trygve O. Tollefsbol, 2007*

370. Adhesion Protein Protocols, *Second Edition,* edited by *Amanda S. Coutts, 2007*

369. Electron Microscopy: *Methods and Protocols, Second Edition,* edited by *John Kuo, 2007*

368. Cryopreservation and Freeze-Drying Protocols, *Second Edition,* edited by *John G. Day and Glyn Stacey, 2007*

367. Mass Spectrometry Data Analysis in Proteomics, edited by *Rune Matthiesen, 2007*

366. Cardiac Gene Expression: *Methods and Protocols,* edited by *Jun Zhang and Gregg Rokosh, 2007*

365. Protein Phosphatase Protocols: edited by *Greg Moorhead, 2007*

364. Macromolecular Crystallography Protocols: *Volume 2, Structure Determination,* edited by *Sylvie Doublié, 2007*

363. Macromolecular Crystallography Protocols: *Volume 1, Preparation and Crystallization of Macromolecules,* edited by *Sylvie Doublié, 2007*

362. Circadian Rhythms: *Methods and Protocols,* edited by *Ezio Rosato, 2007*

361. Target Discovery and Validation Reviews and Protocols: *Emerging Molecular Targets and Treatment Options, Volume 2,* edited by *Mouldy Sioud, 2007*

360. Target Discovery and Validation Reviews and Protocols: *Emerging Strategies for Targets and Biomarker Discovery, Volume 1,* edited by *Mouldy Sioud, 2007*

359. Quantitative Proteomics by Mass Spectrometry, edited by *Salvatore Sechi, 2007*

358. Metabolomics: *Methods and Protocols,* edited by *Wolfram Weckwerth, 2007*

357. Cardiovascular Proteomics: *Methods and Protocols,* edited by *Fernando Vivanco, 2006*

356. High-Content Screening: *A Powerful Approach to Systems Cell Biology and Drug Discovery,* edited by*D. Lansing Taylor, Jeffrey Haskins, and Ken Guiliano, and 2007*

355. Plant Proteomics: *Methods and Protocols,* edited by *Hervé Thiellement, Michel Zivy, Catherine Damerval, and Valerie Mechin, 2007*

354. Plant–Pathogen Interactions: *Methods and Protocols,* edited by *Pamela C. Ronald, 2006*

353. Protocols for Nucleic Acid Analysis by Nonradioactive Probes, *Second Edition,* edited by *Elena Hilario and John Mackay, 2006*

352. Protein Engineering Protocols, edited by *Kristian Müller and Katja Arndt, 2006*

351. *C. elegans:* *Methods and Applications,* edited by *Kevin Strange, 2006*

350. Protein Folding Protocols, edited by *Yawen Bai and Ruth Nussinov 2007*

349. YAC Protocols, *Second Edition,* edited by *Alasdair MacKenzie, 2006*

348. Nuclear Transfer Protocols: *Cell Reprogramming and Transgenesis,* edited by *Paul J. Verma and Alan Trounson, 2006*

347. Glycobiology Protocols, edited by *Inka Brockhausen, 2006*

346. *Dictyostelium discoideum* Protocols, edited by *Ludwig Eichinger and Francisco Rivero, 2006*

345. Diagnostic Bacteriology Protocols, *Second Edition,* edited by *Louise O'Connor, 2006*

344. Agrobacterium Protocols, *Second Edition: Volume 2,* edited by *Kan Wang, 2006*

343. Agrobacterium Protocols, *Second Edition: Volume 1,* edited by *Kan Wang, 2006*

342. MicroRNA Protocols, *edited by Shao-Yao Ying, 2006*

METHODS IN MOLECULAR BIOLOGY™

Serial Analysis of Gene Expression (SAGE)

Methods and Protocols

Edited by

Kåre Lehmann Nielsen

Department of Life Sciences, Aalborg University, Aalborg, Denmark

HUMANA PRESS TOTOWA, NEW JERSEY

Editor
Kåre Lehmann Nielsen
Department of Life Sciences
Aalborg University
Aalborg
Denmark

ISBN: 978-1-58829-676-4 e-ISBN: 978-1-59745-454-4

Library of Congress Control Number: 2007927139

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

All living things carry their genetic information in genes, usually in the form of DNA. The activity of these genes is regulated to meet the requirement by the organism itself or as a response to external abiotic factors such as light, heat, and temperature, but also to biotic factors such as infection by pathogens. Genes are transcribed into mRNAs, which in turn are translated into proteins and catalytically active enzymes. Regulation of this system is primarily obtained by controlling the amount of mRNA that is produced from each gene and the turnover of the corresponding protein. The mRNA population is often referred to as the transcriptome and the protein population as the proteome. The complexity of the system is enormous; all higher organisms, from higher plants to humans, tend to have a similar number of genes, i.e., approx 24,000. In order to understand the genetics that underlie biological change such as development, disease, crop yield, or resistance, it is necessary to perform comparative transcriptomics to understand how the genes are regulated in response to these changes.

Several methods for gene expression profiling exist, such as Northern blotting, Differential Display, EST sequencing, DNA microarrays, and Serial Analysis of Gene Expression (SAGE). The choice of method depends on the need for sensitivity and specificity and whether the methods allow monitoring of genes previously characterized. The dominant method for global gene expression profiling today is DNA microarrays. An array may consist of up to 100,000 unique single-stranded DNA molecules attached to a glass slide in an ordered fashion. An advantage of microarray analysis is that once the array has been made at a high cost, many measurements can be made at a relatively low cost. However, only known genes can be spotted on the array, so it requires a detailed knowledge of the genetic background.

SAGE, on the other hand, can measure the expression of both known and unknown genes. This method relies on the extraction of a unique 14–21 nt sequence (tag) from each mRNA. These tags are ligated together end to end and sequenced. In a typical sequence run of 96 samples approx 1600 tags and, therefore, mRNAs, can be detected. A SAGE study encompasses 50,000 tags and provides detailed knowledge of the 2000 most highly expressed genes in the tissue analyzed. Another application of SAGE is to discover new genes.

Unknown tags obtained through SAGE analysis of a sample can be efficiently used as gene-specific primers in Rapid Amplification of cDNA Ends (RACE) reactions to generate full-length transcripts that can be cloned and sequenced. In principle, a SAGE experiment consists of a series of molecular biology manipulations that can be carried out in any molecular biology laboratory with access to a 96 capillary DNA sequencer. In practice, however, it has proven difficult to achieve enough clones of the appropriate insert length to facilitate efficient detection, and many laboratories have found SAGE a difficult, time consuming, and expensive method.

The aims of *Serial Analysis of Gene Expression (SAGE): Methods and Protocols* are twofold: (i) To enable users, inexperienced with SAGE and having only limited experience in standard molecular biology techniques, to conduct SAGE experiments by providing detailed, commented, tried-and-tested experimental protocols of SAGE and derived methods from experienced researchers across the world. (ii) To facilitate the analysis and comparison of data from SAGE experiments in a correct and efficient way. To achieve this, this book is divided into two parts. Part 1 discusses the experimental procedures of SAGE and related methods such as aRNA-LongSAGE, SuperSAGE, DeepSAGE, and GMAT, and Part 2 discusses the correct extraction and filtering of tags, the analysis of ditag populations, and the performing of statistically correct comparisons of gene expression profiles.

Tag-based gene expression profiling methods, such as SAGE, have been inhibited by the cost of DNA sequencing despite their advantageous global and digital nature. But sequence-based gene expression profiling approaches will become increasingly cost-effective as we approach the $1000 genome with emerging, much cheaper DNA sequencing technologies. It is the hope that *Serial Analysis of Gene Expression (SAGE): Methods and Protocols* may help many laboratories to their first successful experience with tag-based sequencing methods and obtain comprehensive, useful, and interpretable data.

Kåre Lehmann Nielsen

Contents

Contributors

Viatcheslav R. Akmaev • *Bioinformatics, Genzyme Corporation, Framingham, MA*

Frank Baas • *Academic Medical Center, Amsterdam, The Netherlands*

Li Cai • *University of California at Berkeley, Berkeley, CA*

Jeppe Emmersen • *Department of Biochemistry, Chemistry, and Environmental Engineering, University of Aalborg, Aalborg, Denmark*

Xijin Ge • *Evanston Northwestern Healtcare Research Institute, Evanston, IL*

Malali Gowda • *Ohio State University, Columbus, OH*

Anna M. Heidenblut • *Departmental of Internal Medicine, Knappschaftskrankenhaus, Ruhr-University, Bochum, Germany*

Haiyan Huang • *Department of Statistics, University of California at Berkeley, Berkeley, CA*

Annabeth Laursen Høgh • *Department of Biochemistry, Chemistry, and Environmental Engineering, University of Aalborg, Aalborg, Denmark*

Antoine van Kampen • *Academic Medical Center, Amsterdam, The Netherlands*

Marcel Kool • *Academic Medical Center, Amsterdam, The Netherlands*

Detlev H. Krüger • *Humboldt University, Berlin, Germany*

Hideo Matsumura • *Iwate Biotechnology Research Center, Iwate, Japan*

Kåre Lehmann Nielsen • *Department of Life Sciences, University of Aalborg, Aalborg, Denmark*

Monika Reuter • *Humboldt University, Berlin, Germany*

Tae-Young Roh • *National Institute of Health, Bethesda, MD*

Jan M. Ruijter • *Academic Medical Center, Amsterdam, The Netherlands*

Fred van Ruissen • *Department of Neurogenetics, Academic Medical Center, Amsterdam, The Netherlands*

Gerben J. Schaaf • *Department of Human Genetics, Academic Medical Center, Amsterdam, The Netherlands*

Ryohei Terauchi • *Iwate Biotechnology Research Center, Iwate, Japan*

Guo-Liang Wang • *Department of Plant Pathology, Ohio State University, Columbus, OH*

San Ming Wang • *Evanstan Northwestern Healtcare Research Institute, Evanston, IL*

Peter Winter • *University of Frankfurt, Frankfurt am Main, Germany*
Wing Hung Wong • *University of California at Berkeley, Berkeley, CA*
Keji Zhao • *Laboratory of Molecular Immunology, National Institute of Health, Bethesda, MD*

1

Experimental Procedures

1

SAGE and LongSAGE

Annabeth Laursen Høgh and Kåre Lehmann Nielsen

Summary

Serial analysis of gene expression (SAGE) is a high-throughput method for global gene expression analysis that allows the quantitative and simultaneous analysis of a large number of transcripts. SAGE is a digital method and its sensitivity depends only on the number of tags sequenced. Furthermore, SAGE is a powerful tool for finding novel genes that are expressed under certain conditions or in certain tissues. SAGE has been widely used in fields as diverse as cancer research and the development and study of microorganisms. The SAGE method is a series of routine molecular biology procedure and can, at least in principle, be carried out in any laboratory. However, the number of consecutive steps is quite large and in practice, SAGE has been difficult to carry out on a routine basis.

Key Words: Serial analysis of gene expression; SAGE; LongSAGE; global transcriptome profiling.

1. Introduction

Serial analysis of gene expression (SAGE) is a high-throughput method for global gene expression analysis that was introduced by Velculescu et al. in 1995 ***(1)***. SAGE is based on two principles. First, a short nucleotide sequence (tag) from a unique position contains sufficient information to uniquely identify a transcript. Second, these sequence tags can be linked together to form long serial molecules (concatemers) that can be cloned and sequenced ***(1)***. To obtain the tags, mRNA is synthesized into complementary DNA (cDNA) using biotinylated Oligo(dT) **(Fig. 1)**. Double-stranded cDNA is cleaved with a frequent

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

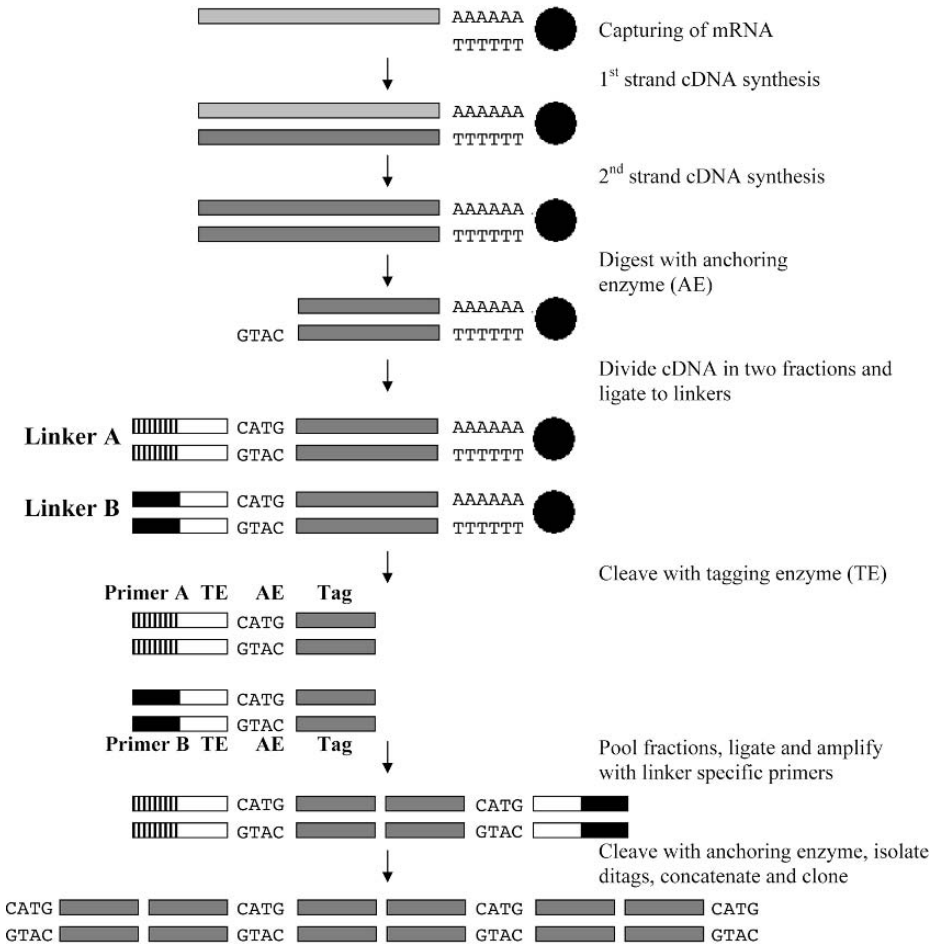


Fig. 1. Schematic overview of serial analysis of gene expression. mRNA is extracted and transcribed into double-stranded complementary (c)DNA on Oligo(dT) streptavidin magnetic beads. cDNA is digested by the anchoring enzyme. The digested cDNA is divided into two fractions, and ligated to different linkers (Linker A and Linker B). The tags are released from the streptavidin magnetic beads by digestion of the tagging enzyme. The linker containing tags are pooled and ligated to form ditags. Following amplification, linkers are removed by digestion of the anchoring enzyme. Ditags are isolated, ligated to form concatemers, cloned, and sequenced.

cutting anchoring enzyme, e.g., *Nla*III, that recognizes the sequence CATG. The 3′-most cDNA fragments are retained using magnetic streptavidin beads. Subsequently, the sample is divided into two fractions, and two different linkers are ligated to the fragments. The linkers contain a restriction site for the tagging enzyme (a type IIS restriction endonuclease, e.g., BsmFI), the anchoring enzyme, and a priming site for PCR. The tagging enzyme cleaves at a defined distance up to 20 bp away from its recognition site, and releases the tags from the magnetic streptavidin beads. The two fractions are pooled, and two sets of linker tag molecules are ligated together to form linker-ditag-linker molecules that can be amplified by PCR using linker-specific primers. Ditags are liberated by digesting with the anchoring enzyme, isolated, and ligated to form concatemers, which are cloned and sequenced ***(1)***. The number of times a particular tag is observed is proportional to the expression level of the corresponding gene. Dinel et al. ***(2)*** have shown that the SAGE method has very good reproducibility, and that the reproducibility, precision, and sensitivity of SAGE are indeed increased by increasing the number of sequenced tags. Furthermore, no *a priori* knowledge of the genes to be identified is required, and the sequence tags can be used to expand sequence information by rapid amplification of cDNA ends (RACE) using cDNA as template ***(3)***.

The original SAGE method was modified into LongSAGE by Saha et al. ***(4)***, generating 21-bp tags instead of 14-bp tags by using another tagging enzyme (*Mme*I instead of BsmFI). The 21-bp tag contains the restriction site of the anchoring enzyme (e.g., CATG) followed by a unique 17-bp tag. In theory, a sequence of 17 bp can distinguish among 17,179,869,184 transcripts (4^{17}) compared to a sequence of 10 bp, which can distinguish among 1,048,576 transcripts (4^{10}). Detailed studies using real sequences show that in practice, SAGE can uniquely identify 94.1 % of *Drosophila melanogaster* genes and 87.6 % of the *Caenorhabditis elegans* genes, whereas LongSAGE uniquely identifies 97.3 % of *D. melanogaster* genes and 93.5 % of the *C. elegans* genes ***(5)***.

2. Materials

2.1. RNA Extraction

1. Liquid nitrogen.
2. Diethylpyrocarbonate (DEPC) water: add 0.75 mL DEPC to 500 mL Milli Q water. Shake well, leave the bottle in a fume cupboard overnight, and autoclave. Store at room temperature.
3. Extraction Buffer: 100 m*M* LiCl, 100 m*M* Tris-HCl pH 8.5, 10 m*M* ethylenediamine tetraacetic acid (EDTA), 1 % sodium dodecyl sulfate (SDS), 15 m*M* dithiothreitol (DTT), in DEPC water.

4. Phenol pH 4.5 (Sigma-Aldrich, St. Louis, MO).
5. Chloroform:isoamyl alcohol (24:1) (Sigma-Aldrich, St. Louis, MO).
6. Phenol:chloroform:isoamyl alcohol (PCI) (25:24:1) (Sigma-Aldrich, St. Louis, MO).

2.2. mRNA Binding to Magnetic Beads

1. Dynabeads Oligo$(dT)_{25}$ (Dynal Biotech Asa, Oslo, Norway).
2. Lysis Buffer: 100 m*M* Tris-HCl, pH 7.5, 500 m*M* LiCl, 10 m*M* EDTA, 1 % lithium dodecyl sulfate, 5 m*M* DTT (Invitrogen, Carlsbad, CA).
3. Wash Buffer A: 10 m*M* Tris-HCl, pH 7.5, 0.15 *M* LiCl, 1 m*M* EDTA, 0.1 % lithium dodecyl sulfate, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
4. Wash Buffer B: 10 m*M* Tris-HCl pH 7.5, 150 m*M* LiCl, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
5. 5X First Strand Buffer: 250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen, Carlsbad, CA).

2.3. cDNA Synthesis

1. DEPC water.
2. dNTP mix, 25 m*M* each (Fermentas, Burlington, Canada).
3. 5X First Strand Buffer: 250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen, Carlsbad, CA).
4. 0.1 *M* DTT (Invitrogen, Carlsbad, CA).
5. SuperScript™ II Reverse Transcriptase (200 U/μL) (Invitrogen, Carlsbad, CA).
6. 5X Second Strand Buffer: 100 m*M* Tris-HCl, pH 6.9, 450 m*M* KCl, 23 m*M* $MgCl_2$, 0.075 m*M* β-NAD^+, 50 m*M* $(NH_4)_2SO_4$ (Invitrogen, Carlsbad, CA).
7. RNase inhibitor (40 U/μL) (New England Biolabs, Ipswich, MA).
8. *Escherichia coli* DNA ligase (10 U/μL) (Invitrogen, Carlsbad, CA).
9. *E. coli* DNA polymerase (10 U/μL) (Invitrogen, Carlsbad, CA).
10. *E. coli* RNase H (5 U/μL) (Fermentas, Burlington, Canada).
11. 0.5 *M* EDTA (Bie & Berntsen A-S, Rødovre, Denmark).
12. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
13. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL bovine serum albumin (BSA) (New England Biolabs, Ipswich, MA).
14. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).

2.4. Cleavage of cDNA With the Anchoring Enzyme NlaIII

1. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
2. 100X BSA (New England Biolabs, Ipswich, MA).
3. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).

4. *Nla*III (10 U/μL) (New England Biolabs, Ipswich, MA).
5. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
6. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL BSA (New England Biolabs, Ipswich, MA).

2.5. Ligating Linkers to Bound cDNA

1. Linker1 A (TAG Copenhagen, Denmark):
   5′-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA TCC GAC ATG-3′
   Linker1 B (TAG Copenhagen, Denmark):
   5′-PO_4 TCG GAT ATT AAG CCT AGT TGT ACT GCA CCA GCA AAT CC (Amino C7)-3′
   Linker2 A (TAG Copenhagen, Denmark):
   5′-TTT CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC GTC CGA CAT G-3′
   Linker2 B (TAG Copenhagen, Denmark):
   5′-PO_4 TCG GAC GTA CAT CGT TAG AAG CTT GAA TTC GAG CAG (Amino C7)-3′
   Linker oligonucleotides (TAG Copenhagen, Denmark) are dissolved in DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany) to a final concentration of 100 μ*M* and stored at −20 °C.
2. 10X T4 DNA Ligase Buffer: 400 m*M* Tris-HCl, 100 m*M* $MgCl_2$, 100 m*M* DTT, 5 m*M* ATP (Fermentas, Burlington, Canada).
3. DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany).
4. LS Adaptor A (Linker1 A + Linker1 B) (60 ng/μL).
5. LS Adaptor B (Linker2 A + Linker2 B) (60 ng/μL).
6. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH7.5.
7. T4 DNA Ligase (5 U/μL) (Fermentas, Burlington, Canada).
8. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL BSA (New England Biolabs, Ipswich, MA).

2.6. Release of cDNA Tags Using the Tagging Enzyme *Mme*I

1. 32 m*M* *S*-adenosylmethionine (SAM) (New England Biolabs, Ipswich, MA).
2. DEPC water.
3. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).
4. DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany).
5. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA.
6. *Mme*I (2 U/mL) (New England Biolabs, Ipswich, MA).

7. PCI (25:24:1) (Sigma-Aldrich, St. Louis, MO).
8. 7.5 *M* ammonium acetate (Sigma-Aldrich, St. Louis, MO).
9. Glycogen (20 mg/mL) (Fermentas, Burlington, Canada).
10. 100 % ethanol (Sigma-Aldrich, St. Louis, MO).
11. 70 % ethanol.

2.7. Ligating Tags to Form Ditags

1. 3 m*M* Tris-HCl, pH 7.5.
2. 10X T4 DNA Ligase Buffer (Fermentas, Burlington, Canada).
3. DEPC water.
4. T4 DNA Ligase (5 U/μL) (Fermentas, Burlington, Canada).
5. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH7.5.

2.8. PCR Amplification of Ditags

1. Cold DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany).
2. 10X PCR Buffer: 100 m*M* Tris-HCl, pH 8.3, 500 m*M* KCl, 15 m*M* $MgCl_2$, 1 % Triton X-100 (Bie & Berntsen A-S, Rødovre, Denmark).
3. Primer mix (LS DTP 1+2)
   LS DTP-1 primer: 5′-Biotin GTG CTC GTG GGA TTT GCT GGT GCA GTA CA-3′
   LS DTP-2 primer: 5′-Biotin GAC CTC GTG CTG CTC GAA TTC AAG CTT CT-3′
   Primer oligonucleotides (MWG-Biotec AG, Ebersberg, Germany) are dissolved in DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany) to a final concentration of 100 pmol/μL and stored at −40 °C.
4. 25 m*M* $MgCl_2$ (Fermentas, Burlington, Canada).
5. dNTP mix, 25 m*M* each (Fermentas, Burlington, Canada).
6. Taq polymerase (5 U/μL) (Bie & Berntsen A-S, Rødovre, Denmark).
7. 30 % acrylamide/bis (19:1) (AppliChem, Darmstadt, Germany).
8. Sterile, autoclaved MilliQ water.
9. 50X TAE Buffer: 2 *M* Tris, 1 *M* acetic acid, 0.05 *M* EDTA, pH 8.3.
10. *N,N,N',N'*-tetramethylethylendiamine (TEMED) (Bie & Berntsen A-S, Rødovre, Denmark).
11. Ammonium persulfate: prepare 10 % solution in water and store at −20 °C.
12. Molecular weight markers for gel electrophoresis: GeneRuler™ 100-bp DNA ladder (Fermentas, Burlington, Canada) and 25-bp DNA ladder (Invitrogen, Carlsbad, CA), diluted to a final concentration of approx 0.1 μg/μL.
13. Running Buffer (1X TAE Buffer): 40 m*M* Tris, 20 m*M* acetic acid, 1 m*M* EDTA pH 8.3.
14. 6X TAE Loading Buffer: 240 m*M* Tris, 120 m*M* acetic acid, 6 m*M* EDTA pH 8.3, 17 % glycerol, bromophenol blue.

15. Staining solution: 25 mL 1X TAE containing 5 μL ethidium bromide 10 mg/mL (Sigma-Aldrich, St. Louis, MO).
16. PCI (25:24:1) (Sigma-Aldrich, St. Louis, MO).
17. 7.5 *M* ammonium acetate (Sigma-Aldrich, St. Louis, MO).
18. Glycogen 20 mg/mL (Fermentas Burlington, Canada).
19. 100 % ethanol (Sigma-Aldrich, St. Louis, MO).
20. 70 % ethanol.
21. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
22. TE: 10 m*M* Tris-HCl, pH 7.5, 1 m*M* EDTA, pH 7.5.
23. 50-mL polypropylene centrifuge tubes (Sorvall).
24. 10-mL glass pipet for PCI extraction (*see* **Note 1**).

2.9. Isolation of Ditags

1. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).
2. 100X BSA (New England Biolabs, Ipswich, MA).
3. *Nla*III (10 U/μL) (New England Biolabs, Ipswich, MA).
4. 30 % acrylamide/bis (19:1) (AppliChem, Darmstadt, Germany).
5. Sterile, autoclaved MilliQ water.
6. 50X TAE Buffer: 2 *M* Tris, 1 *M* acetic acid, 0.05 *M* EDTA, pH 8.3.
7. TEMED (Bie & Berntsen A-S, Rødovre, Denmark).
8. Ammonium persulfate: prepare 10 % solution in water and store at −20 °C.
9. Molecular weight markers for gel electrophoresis: GeneRuler 100-bp DNA ladder (Fermentas, Burlington, Canada) and 25-bp DNA ladder (Invitrogen, Carlsbad, CA), diluted to a final concentration of approx 0.1 μg/μL.
10. Running Buffer (1X TAE Buffer): 40 m*M* Tris, 20 m*M* acetic acid, 1 m*M* EDTA pH 8.3
11. 6X TAE Loading Buffer: 240 m*M* Tris, 120 m*M* acetic acid, 6 m*M* EDTA pH 8.3, 17 % glycerol, bromophenol blue.
12. Staining solution: 25 mL 1X TAE containing 5 μL ethidium bromide 10 mg/mL (Sigma-Aldrich, St. Louis, MO).
13. PCI (25:24:1) (Sigma-Aldrich, St. Louis, MO).
14. 7.5 *M* ammonium acetate (Sigma-Aldrich, St. Louis, MO).
15. Glycogen 20 mg/mL (Fermentas, Burlington, Canada).
16. 100 % ethanol (Sigma-Aldrich, St. Louis, MO).
17. 70 % ethanol.
18. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
19. Spin-X® tubes (Corning Costar Inc., NY).

2.10. Concatenation of Ditags

1. DNA synthesis and protein sequencing grade water (AppliChem, Darmstadt, Germany).

2. 10X T4 DNA Ligase Buffer: 400 m*M* Tris-HCl, 100 m*M* $MgCl_2$, 100 m*M* DTT, 5 m*M* ATP (Fermentas, Burlington, Canada).
3. T4 Ligase (5U/μL) (Fermentas, Burlington, Canada).
4. *Nla*III (10U/μL) (New England Biolabs, Ipswich, MA).
5. 30 % acrylamide/bis (19:1) (AppliChem, Darmstadt, Germany).
6. Sterile, autoclaved MilliQ water.
7. 50X TAE Buffer: 2 *M* Tris, 1 *M* acetic acid, 0.05 *M* EDTA, pH 8.3.
8. TEMED (Bie & Berntsen A-S, Rødovre, Denmark).
9. Ammonium persulfate: prepare 10 % solution in water and store at −20 °C.
10. Molecular weight markers for gel electrophoresis: GeneRuler 100-bp DNA ladder (Fermentas, Burlington, Canada) and 25-bp DNA ladder (Invitrogen, Carlsbad, CA), diluted to a final concentration of approx 0.1 μg/μL.
11. Running Buffer (1X TAE Buffer): 40 m*M* Tris, 20 m*M* acetic acid, 1 m*M* EDTA pH 8.3.
12. 6X TAE Loading Buffer: 240 m*M* Tris, 120 m*M* acetic acid, 6 m*M* EDTA pH 8.3, 17 % glycerol, bromophenol blue.
13. Staining solution: 25 mL 1X TAE containing 5 μL ethidium bromide 10 mg/mL (Sigma-Aldrich, St. Louis, MO).
14. Electroelution device (Elutrap system from Schleicher & Schuell, Dassel, Germany).
15. PCI (25:24:1) (Sigma-Aldrich, St. Louis, MO).
16. 7.5 *M* ammonium acetate (Sigma-Aldrich, St. Louis, MO).
17. Glycogen 20 mg/mL (Fermentas, Burlington, Canada).
18. 100 % ethanol (Sigma-Aldrich, St. Louis, MO).
19. 70 % ethanol.
20. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.

2.11. Cloning of Concatemers

1. pZErO-1 (1 μg/μL, part of the Zero Background Cloning kit) (Invitrogen Carlsbad, CA).
2. 10X Buffer 2: 100 m*M* Tris-HCl, pH 7.9, 10 m*M* $MgCl_2$, 50 m*M* NaCl, 1 m*M* DTT (New England Biolabs, Ipswich, MA).
3. *Sph*I (5 U/μL) (New England Biolabs, Ipswich, MA).
4. Sterile, autoclaved MilliQ water.
5. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
6. PCI (25:24:1) (Sigma-Aldrich, St. Louis, MO).
7. 7.5 *M* ammonium acetate (Sigma-Aldrich, St. Louis, MO).
8. Glycogen 20 mg/mL (Fermentas, Burlington, Canada).
9. 100 % ethanol (Sigma-Aldrich, St. Louis, MO).
10. 70 % ethanol.
11. 10X T4 DNA Ligase Buffer: 400 m*M* Tris-HCl, 100 m*M* $MgCl_2$, 100 m*M* DTT, 5 m*M* ATP (Fermentas, Burlington, Canada).
12. T4 DNA Ligase (5 U/μL) (Fermentas, Burlington, Canada).

2.12. Transformation of Escherichia coli

1. Positive control: Supercoiled pUC19 plasmid (Invitrogen, Carlsbad, CA).
2. TOP10 Electrocompetent *E. coli* (Invitrogen, Carlsbad, CA).
3. Electroporation device (EasyjecT Prima, EQUIBIO, Oxford, UK).
4. 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 (Invitrogen, Carlsbad, CA).
5. Low-salt Luria-Bertani (LB) medium: 1 % (w/v) tryptone or casein peptone, 0.5 % (w/v) yeast extract, 0.5 % (w/v) NaCl, pH 7.5.
6. Zeocin (100 mg/mL, store at −20 °C in darkness, this antibiotic is light-sensitive) (Invitrogen, Carlsbad, CA).
7. 500 m*M* IPTG (AppliChem, Darmstadt, Germany).
8. X-gal: 40 mg dissolved in 200 µL dimethylsulfoxide (DMSO) (Fermentas, Burlington, Canada; Sigma-Aldrich, St. Louis, MO).
9. Low-salt LB agar with Zeocin, X-gal and IPTG: 50 µg/mL Zeocin, 80 µg/mL X-gal, 1 m*M* IPTG, 1.5 % (w/v) agar (Bie & Berntsen A-S, Rødovre, Denmark).
10. Sterile 96-well culture plates.
11. Autoclaved toothpicks.

3. Methods

The LongSAGE method described in this chapter is essentially based on the technique developed by Velculescu and coworkers for obtaining a digital genome-wide expression profile of the genes involved in selected tissues ***(1**,**3**, **4**,**6)*** The LongSAGE technique generates a unique 17- to 21-bp tag for each transcript, as opposed to the 10- to 14-bp tag generated by the original SAGE method.

3.1. RNA Extraction From Potato Tuber Tissue

1. Place Phenol and Extraction Buffer in water bath at 60 °C.
2. Grind 5.0 g of frozen tissue in a mortar with liquid N_2 (*see* **Note 2**).
3. Transfer frozen powder to a 40-mL (polypropylene) centrifuge tube with a screw cap.
4. Add 10 mL warm Extraction Buffer + 10 mL warm Phenol. Put on screw cap and shake vigorously for 30 s.
5. Add 10 mL chloroform:isoamyl alcohol (24:1). Shake vigorously for 30 s.
6. Centrifuge for 10 min at 9682*g* and 4 °C.
7. Transfer upper aqueous phase to a fresh 40-mL centrifuge tube, and add 10 mL of PCI (25:24:1). Shake for 30 s.
8. Centrifuge for 10 min at 9682*g* and 4 °C.
9. Transfer upper aqueous phase to a fresh 40-mL centrifuge tube and add 10 mL of PCI (25:24:1). Shake for 30 s.
10. Centrifuge for 10 min at 9682*g* and 4 °C.

11. Transfer upper aqueous phase to a fresh 40-mL centrifuge tube and add 10 mL of chloroform:isoamyl alcohol (24:1). Shake for 30 s.
12. Centrifuge for 10 min at 9682*g* and 4 °C.
13. Transfer aqueous phase to a fresh 40-mL centrifuge tube and add 10 mL of chloroform:isoamyl alcohol (24:1). Shake for 30 s.
14. Centrifuge for 10 min at 9682*g* and 4 °C.
15. Transfer upper aqueous phase to a fresh 40-mL centrifuge tube. Measure the volume with sterile pipet. Add equal volume of 4 *M* LiCl. Mix gently, and store tube overnight at 4 °C.
16. The following day, centrifuge for 40 min at 9682*g* and 4 °C.
17. Decant supernatant (carefully) and immediately add 10 mL of chilled 70 % ethanol to rinse pellet. Carefully decant ethanol and air-dry pellet for approx 15 min at room temperature.
18. Resuspend pellet in 150 µL DEPC water, and transfer to a sterile 1.5-mL microcentrifuge tube. Take out a sample of 5–10 µL for determination of total RNA concentration by absorption at 260 nm.
19. Freeze the rest of the sample at −40 °C.

3.2. mRNA Binding to Magnetic Beads

1. Thoroughly resuspend the Oligo(dT)$_{25}$ beads and transfer 100 µL to an RNase-free, 1.5-mL tube.
2. Place the tube on a magnetic stand for 1–2 min, and discard supernatant.
3. Wash Oligo(dT)$_{25}$ beads by resuspending them in 500 µL of Lysis Buffer. Place the tube in a magnetic stand.
4. Prepare RNA sample for binding to the beads. Use 5–50 µg of total RNA or 50–100 ng of mRNA (*see* **Note 3**). Adjust volume to 1 mL with Lysis Buffer.
5. Carefully remove the supernatant from Oligo(dT)$_{25}$ beads, and immediately add the RNA sample.
6. Mix Oligo(dT)$_{25}$ beads and RNA sample by slowly rocking the tube on a rocking platform for 30 min at room temperature and 50 rpm.
7. Place the tube on a magnetic stand for 1–2 min, and carefully remove supernatant.
8. Wash the beads twice by placing the tube on magnetic stand for 1–2 min and removing the supernatant between washes with 1 mL Wash Buffer A.
9. Wash with 1 mL Wash Buffer B.
10. Wash the Oligo(dT)$_{25}$ beads four times with 100 µL 1X First Strand Buffer by placing the tube on magnetic stand for 1–2 min and removing the supernatant between washes. After the fourth wash, do not remove supernatant (*see* **Note 4**).

3.3. cDNA Synthesis

1. Prepare the first strand cDNA reaction mix by mixing the following reagents on ice: 18 µL 5X First Strand Buffer, 1 µL RNase inhibitor, 57.2 µLDEPC water, 9 µL 0.1 *M* DTT, 1.8 µL dNTP mix.

2. Remove the supernatant from the Oligo$(dT)_{25}$ beads and resuspend Oligo$(dT)_{25}$ beads containing mRNA in the first strand cDNA reaction mix. Mix gently by flicking the tube with a finger. Store the tube at 37 °C for 2 min to equilibrate the reagents.
3. Add 3 μL reverse transcriptase. Mix gently and incubate at 37–42 °C for 1 h. Mix gently at every 10–15 min by flicking the tube.
4. Meanwhile, equilibrate another water bath to 16 °C for second strand synthesis.
5. Chill the first strand reaction on ice for 2 min, and add the following second strand reagents in the following order to the tube containing 90 μL of the first strand reaction: 474 μL DEPC water, 150 μL 5X Second Strand Buffer, 6 μL dNTP mix, 5 μL *E. coli* DNA ligase, 20 μL *E. coli* DNA polymerase, 2 μL *E. coli* RNase H. Mix contents by vortexing and centrifuge the tube briefly in a benchtop centrifuge. Incubate the reaction mixture at 16 °C for 2 h, mixing gently every 10–15 min to resuspend the Oligo$(dT)_{25}$ beads.
6. During incubation, preheat Wash Buffer C to 75 °C.
7. Place the reaction tube on ice and add 45 μL 0.5 *M* EDTA to stop the reaction. Place the tube on a magnetic stand for 1–2 min and carefully remove supernatant (*see* **Note 5**). Add 750 μL warm Wash Buffer C to inactivate the *E. coli* DNA polymerase. Mix well and heat the sample to 75 °C for 10–12 min with intermittent mixing to completely inactivate the polymerase. Place the tube on magnetic stand for 1–2 min and remove supernatant. Wash again with 750 μL Wash Buffer C. Perform the wash quickly to prevent precipitation of SDS, which may trap the beads.
8. Wash sample three times with 750 μL Wash Buffer D and then resuspend beads in 750 μL Wash Buffer D. (Remove 5 μL of sample to determine the efficiency of the cDNA synthesis reaction. Store as sample 1 [S1] at 4 °C; *see* **Fig. 2**.)
9. Place the tube on magnetic stand for 1–2 min and carefully remove supernatant.
10. Add 200 μL 1X Buffer 4 to the tube and gently resuspend Oligo$(dT)_{25}$ beads. Transfer the contents of the tube to a new tube to avoid any traces of exonuclease activity from *E. coli* DNA polymerase. Wash the old tube with 200 μL 1X Buffer 4 and transfer the contents to the new tube containing the reaction mix.
11. Place tube on magnetic stand for 1–2 min and remove supernatant.
12. Wash Oligo$(dT)_{25}$ beads once with 200 μL 1X Buffer 4.

3.4. Cleavage of cDNA With the Anchoring Enzyme NlaIII

1. Remove supernatant and resuspend Oligo$(dT)_{25}$ beads in: 172 μL LoTE, 2 μL 100X BSA, 20 μL 10X Buffer 4, 6 μL *Nla*III. Incubate for 2.5 h (or 1 h) at 37 °C. Mix occasionally by flicking the tube with a finger.
2. Meanwhile, heat Wash Buffer C to 37 °C to prevent SDS precipitation.
3. After the reaction is complete, place the tube containing Oligo$(dT)_{25}$ beads on a magnetic stand for 1–2 min and carefully discard supernatant.

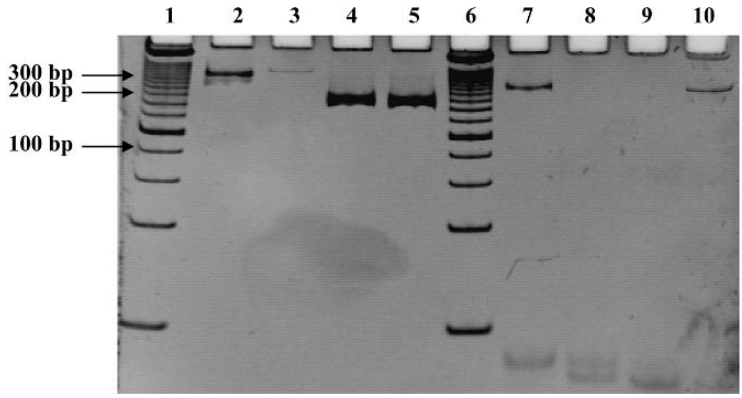


Fig. 2. Verification of complementary (c)DNA synthesis, *Nla*III digestion and ligation of linkers. **1:** 25-bp DNA ladder. **2:** cDNA template, *Solanum tuberosum* p23S and p23AS primers used. A band should appear at 314 bp. **3:** *Nla*III template, *S. tuberosum* p23S and p23AS primers used. *Nla*III digestion occurs between the two primer sites and therefore no band should appear. **4:** cDNA template, *S. tuberosum* 60SL3S and 60SL3AS primers used. A band should appear at 201 bp. **5:** *Nla*III template, *S. tuberosum* 60SL3S and 60SL3AS primers used. A band should appear at 201 bp. **6:** 25-bp DNA ladder. **7:** LS Adapter A template, LS DTP-1 and *S. tuberosum* 60SL3AS primers used. A band should appear at 246 bp. **8:** LS Adapter B template, LS DTP-1, and *S. tuberosum* 60SL3AS primers used. No band is seen with unmatched primers and adaptors. **9:** LS Adapter A template, LS DTP-2 and *S. tuberosum* 60SL3AS primers used. No band is seen with unmatched primers and adaptors. **10:** LS Adapter B template, LS DTP-2 and *S. tuberosum* 60SL3AS primers used. A band should appear at 244 bp.

4. Inactivate *Nla*III by washing the tube twice with 750 µL warm Wash Buffer C. Wash sample three times with 750 µL Wash Buffer D. (Remove 5 µL of sample to determine the efficiency of the *Nla*III digestion. Store as sample 2 [S2] at 4 °C; *see* **Fig. 2**.)
5. Proceed to Ligating linkers to bound cDNA or store tube at 4 °C overnight.

3.5. Ligating Linkers to Bound cDNA

Prior to use, the linker oligonucleotides are phosphorylated and hybridized to obtain LS Adaptor A and LS Adaptor B. Phosphorylated and hybridized linkers can be stored at −20 °C in aliquots for single use.

1. Mix the following in two separate tubes; LS Adaptor A: 17 µL Linker1 A, 17 µL Linker1 B, 4 µL 10X Buffer ATP Poly Nucleotide Kinase; LS Adaptor B: 17 µL Linker2 A, 17 µL Linker2 B, 4 µL 10X Buffer ATP Poly Nucleotide Kinase.

2. Place tubes in 100 mL boiling water, and incubate for 1 h leaving water container at room temperature.
3. Load LS Adaptor ligation mix onto a 10 × 8 cm × 0.75 mm 15 % TAE-polyacrylamide gel (comb with 10 lanes). Add 10 μL 6X TAE loading buffer to each tube. Load samples and 25-bp DNA ladder on gel. Conduct gel electrophoresis in 1X TAE running buffer at 120 V for 1 h. Stain gel in 25 mL 1X TAE running buffer containing 5 μL ethidium bromide for 5 min at room temperature.
4. Isolate 40-bp bands for LS Adaptors A and B using a scalpel. Transfer gel pieces to two 0.5-mL Eppendorf tubes marked AdA and AdB, respectively. These tubes contain a hole made by a sterile needle. Place the 0.5-mL tubes in a 1.5-mL Eppendorf tube marked AdA and AdB, respectively.
5. Centrifuge at maximum speed in a benchtop centrifuge for 30–60 s.
6. Add 500 μL of the following elution buffer to each tube: 800 μL LoTE, 200 μL 7.5 *M* ammonium acetate. Elute adaptors form the gel pieces overnight at 4 °C or at 37 °C for 45 min. Transfer content to 2-mL Spin-X filter tubes and centrifuge at maximum speed in a benchtop centrifuge for approx 30 s. Add 2 μL glycogen and 1500 μL 100 % ethanol to each tube and vortex briefly. Store the tube at −40 °C for 30 min, centrifuge at maximum speed in a benchtop centrifuge for 30 min, and remove and discard supernatant. Wash pellets in 1 mL 70 % ethanol. Leave the tube on the bench top with the lid open and air-dry pellets for a minimum of 10 min.
7. Resuspend each pellet in 100 μL TE and determine LS Adaptor concentrations.
8. To ligate adaptors to immobilized cDNA, place the tube with beads on a magnetic stand for 1–2 min and carefully remove supernatant.
9. Wash Oligo$(dT)_{25}$ beads twice with 150 μL of 1X T4 DNA Ligase Buffer. Immediately after the final wash, resuspend beads in 100 μL 1X T4 DNA Ligase Buffer and divide the sample into two new tubes labeled A and B. Be careful to divide the Oligo$(dT)_{25}$ beads while they are resuspended, as the Oligo$(dT)_{25}$ beads may stick to the original tube or pipet tips.
10. Wash each tube (A and B) once with 50 μL 1X T4 DNA Ligase Buffer. Resuspend Oligo$(dT)_{25}$ beads in 50 μL 1X T4 DNA Ligase Buffer.
11. Place tubes (A and B) on a magnetic stand for 1–2 min and carefully remove supernatant. Transfer the tubes to ice and add the following reagent to the beads: Tube A, 1 μL LS Adaptor A (60 ng/μL), 14.5 μL LoTE, 2 μL 10X T4 DNA Ligase Buffer; Tube B, 1 μL LS Adaptor B (60 ng/μL), 14.5 μL LoTE, 2 μL 10X T4 DNA Ligase Buffer.
12. Resuspend Oligo$(dT)_{25}$ beads by flicking each tube. Heat the tube for at least 2 min at 50 °C. Cool the tube for 15 min at room temperature and then chill on ice. Add 2 μL T4 DNA Ligase to each tube and mix well. Incubate overnight at 16 °C.
13. The following day, wash each tube three times with 500 μL of Wash Buffer D. (Remove 5 μL of resuspended Oligo$[dT]_{25}$ beads from each tube to determine the efficiency of the adapter ligation. Store samples 3A and 3B [S3A and S3B] at 4 °C; *see* **Fig. 2**.)

3.6. Release of cDNA Tags Using the Tagging Enzyme MmeI

1. Prepare 10X SAM by adding 1 μL 32 m*M* SAM to 79 μL DEPC water.
2. Prepare 1X Buffer 4/1X SAM by combining 80 μL 10X Buffer 4, 720 μL DNA synthesis and protein sequencing grade water, 1 μL 32 m*M* SAM.
3. Place each tube (A and B) on a magnetic stand for 2 min and remove supernatant.
4. Wash each tube twice with 200 μL 1X Buffer 4/1X SAM. Carefully remove and discard the supernatant and place tubes on ice.
5. Add the following to each tube: 70 μL LoTE, 10 μL 10X Buffer 4, 10 μL 10X SAM, 10 μL *Mme*I. Incubate tubes at 37 °C for 2.5 h with occasional gentle mixing.
6. Place tubes on magnetic stand for 2 min. Do not discard the supernatant (*see* **Note 6**). Carefully remove the supernatant from each tube and pool supernatants from each tube to a new tube labeled A + B. Add 100 μL of LoTE to tube (A + B) to yield a total volume of 300 μL.
7. Add 300 μL PCI (25:24:1) to tube and vortex thoroughly. Centrifuge for 5 min at room temperature and at maximum speed.
8. Transfer 300 μL of the upper aqueous phase to a new tube. Remove 200 μL from this tube to a new tube (A + B). The remaining 100 μL are used as a negative control (no ligase). Add 100 μL DEPC water to tube (no ligase) to yield a final volume of 200 μL.
9. To each tube (200 μL of sample and 200 μL of negative control), add 133 μL 7.5 *M* ammonium acetate, 3 μL glycogen and 1 mL of 100 % ethanol. Mix virgorously.
10. Store tubes at −40 °C for a minimum of 30 min and centrifuge at maximum speed in a benchtop centrifuge for 30–40 min at 4 °C.
11. Carefully remove supernatant from each tube and discard it. Be careful not to disturb the pellet.
12. Wash each pellet twice with 1 mL of cold 70 % ethanol. After the final wash, centrifuge each tube again to collect any residual ethanol. Carefully remove the ethanol by pipet and air-dry for 5–10 min.
13. Resuspend the sample (A + B) pellet in 4 μL LoTE and the negative control (no ligase) pellet in 2 μL LoTE and incubate at 37 °C for 10–15 min to aid insolubilization.

3.7. Ligating Tags to Form Ditags

1. Prepare 2X ditag reaction in a sterile microcentrifuge tube on ice by combining: 1.5 μL 3 m*M* Tris-HCl, pH 7.5, 0.9 μL 10X T4 DNA Ligase Buffer, 1.1 μL DEPC water, 1 μL T4 DNA Ligase.
2. Prepare 2X Negative Control in a sterile microcentrifuge tube on ice by combining: 2.25 μL 3 m*M* Tris-HCl, pH 7.5, 0.75 μL 10X T4 DNA Ligase Buffer, 0.75 μL DEPC water.
3. Add 4 μL of 2X Ditag Reaction Mix to the tags resuspended in 4 μL LoTE (A + B).
4. Add 2 μL of 2X Negative Control Mix to the negative control (no ligase)
5. Incubate tubes overnight at 16 °C.

3.8. PCR Amplification of Ditags

To optimize PCR conditions, a test PCR is performed using different dilutions of ditags (1:40, 1:80, 1:160, 1:320, 1:640 in DNA synthesis and protein sequencing grade water). Ligated LS Adaptors are used as positive control (1:40). Two PCR reactions containing the negative control (no ligase) as template, and no template are used as negative controls, respectively (*see* **Fig. 3**).

1. For each PCR reaction mix the following on ice: 1 μL template (diluted ditags and controls), 36.5 μL cold DNA synthesis and protein sequencing grade water, 5 μL 10X PCR Buffer, 1 μL LS DTP primer mix, 5 μL $MgCl_2$, 1 μL dNTP mix, 0.5 μL Taq polymerase (*see* **Note 7**).
2. Perform PCR according to the following procedure: initial denaturation at 94 °C for 1 min, followed by 28 cycles of denaturation at 94 °C for 30 s, annealing at 53.5 °C for 1 min, and elongation at 70 °C for 1 min (*see* **Note 8**).
3. Load 5 μL of each PCR reaction onto a 10 × 8 cm 0.75 mm 15 % TAE-polyacrylamide gel (comb with 10 lanes). Conduct gel electrophoresis in 1X TAE running buffer for 45 min at 100 V and 45 min at 120 V. Stain the gel for 5 min in 25 mL 1X TAE containing 5 μL ethidium bromide (*see* **Fig. 3**).
4. Select the most appropriate ditag dilution as template for PCR amplification.

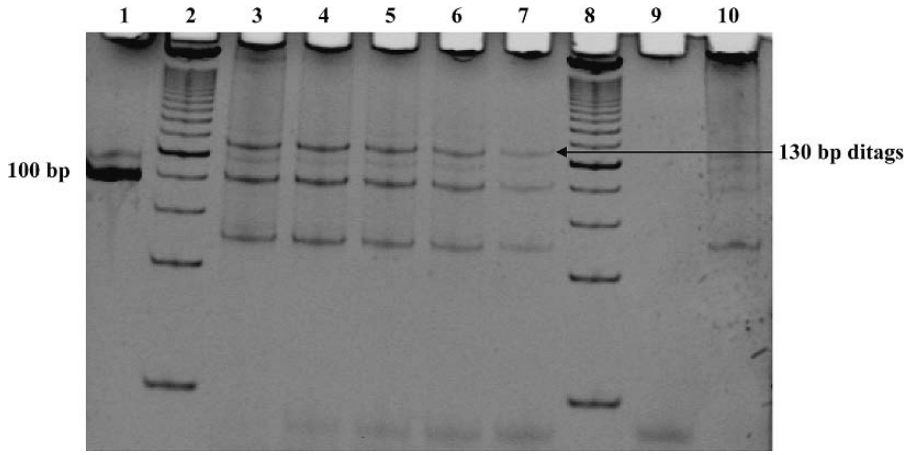


Fig. 3. Ditag PCR. **1,** 1:40 positive control (Linker A + B); **2,** 25-bp DNA ladder; **3,** 1:40 template; **4,** 1:80 template; **5,** 1:160 template; **6,** 1:320 template; **7,** 1:640 template; **8,** 25-bp DNA ladder; **9,** negative control (no template); **10,** 1:20 negative control (no ligase). The ditag bands are located at approx 130 bp. The bands seen at 100 bp correspond to ligated linkers without any ditag.

5. Make 200 50-μL PCR reactions. Mix the PCR reactions in a 50-mL greiner Blue Cap tube covered with ice. Add the reagents in the following order: 7.3 mL cold DNA synthesis and protein sequencing grade water, 1 mL 10X PCR Buffer, 200 μL LS DTP primer mix, 1 mL $MgCl_2$, 200 μL dNTP mix, 100 μL Taq polymerase and 200 μL template (*see* **Note 9**).
6. Mix PCR reaction mix gently by inverting the tube twice.
7. Add 50 μL PCR reaction mix to each well of two 96-well PCR plates on ice using a multichannel pipet.
8. Perform PCR according to the following procedure: initial denaturation at 94 °C for 1 min, followed by 28 cycles of denaturation at 94 °C for 30 s, annealing at 53.5 °C for 1 min, and elongation at 70 °C for 1 min (*see* **Note 8**).
9. Pool PCR reactions and place on ice. Analyze 5 μL and 10 μL on a 10 × 8 cm × 0.75 mm 15 % TAE-polyacrylamide gel (comb with 10 lanes). Add 1 μL 6X TAE loading buffer per 5 μL PCR sample. Load samples and 25-bp DNA ladder on gel. Conduct gel electrophoresis in 1X TAE running buffer for 45 min at 100 V and 45 min at 120 V. Stain the gel for 5 min in 25 ml 1X TAE containing 5 μL ethidium bromide. Visualize bands by exposure to ultraviolet (UV) light (*see* **Fig. 4**).
10. Store PCR reactions at −40 °C or proceed to isolation of ditags.

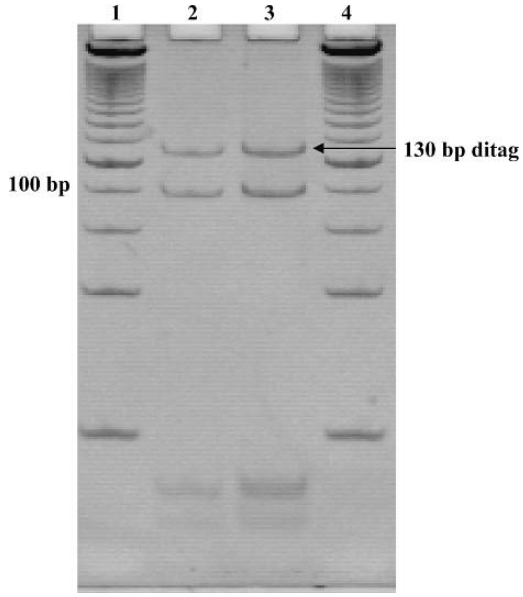


Fig. 4. PCR of ditags. **1,** 25-bp DNA ladder; **2,** 5-μL PCR sample; **3,** 10-μL PCR sample; **4,** 25-bp DNA ladder. The ditag bands are located at approx 130 bp. The bands seen at 100 bp correspond to ligated linkers without any ditag.

3.9. Isolation of Ditags

1. Extract PCR products in polypropylene tubes by adding an equal amount of PCI to PCR products (*see* **Note 10**). Vortex well. Centrifuge at 2400*g* for 10 min at room temperature and transfer the upper aqueous phase (~9 mL) to a new tube (*see* **Note 11**).
2. Divide sample (4.5 mL) into two centrifuge tubes, and add the following to each tube: 1150 µL 7.5 *M* ammonium acetate, 26 µL glycogen, 12.25 mL cold 100 % ethanol (*see* **Note 12**).
3. Mix vigorously. Store tubes at −40 °C for a minimum of 30 min or overnight. Centrifuge at 12,000*g* for 30 min at 4 °C. Carefully remove and discard supernatant.
4. Wash pellet twice with approx 25 mL cold 70 % ethanol. Carefully remove ethanol and air-dry pellet for 15–20 min (*see* **Note 13**).
5. Resuspend pellets in 2X 250 µL LoTE, and incubate tubes at 37 °C for 5–10 min to aid in solubilization. Centrifuge for 5 min at maximum speed. Transfer supernatant to a new tube.
6. Digest with *Nla*III (*see* **Note 14**) by adding the following to each 250-µL sample: 30 µL 10X Buffer 4, 3 µL 100X BSA, 10 µL *Nla*III. Mix contents well and incubate at 37 °C for 2 h.
7. Add 60 µL of 6X TAE loading buffer to approx 300 µL of sample from *Nla*III digestion. Load 360 µL onto a 10 × 8 cm × 1.5 mm 15 % TAE-polyacrylamide gel.
8. Conduct gel electrophoresis in 1X TAE running buffer at 90 V for 45 min and 100 for 50 min. Stain the gel for 5 min in 25 mL 1X TAE containing 5 µL ethidium bromide. Visualize bands by exposure to UV light. Excise the 34-bp product using a clean scalpel (*see* **Fig. 5**).
9. Electroelute ditags by preparing the electroelution device in such a way that the trap is 1 U-insert wide, place the gel slice into the elution chamber and electroelute

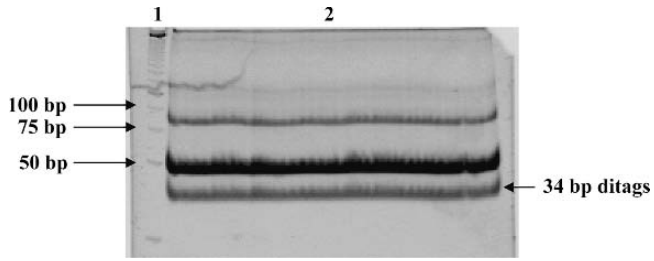


Fig. 5. Isolation of 34-bp ditags. **1,** 25-bp DNA ladder; **2,** *Nla*III-digested 130-bp ditag sample. The band seen between 25 bp and 50 bp corresponds to 34 bp ditag. The band at 50 bp corresponds to linkers. The band between 75 bp and 100 bp corresponds to ditags with one of the linkers. The band at approx 130 bp correspond to undigested 130-bp ditags.

in 1X TAE running buffer at 100 V and room temperature for 1 h 15 min. Remove buffer from trap (200–250 µL) and put into a new tube.

10. Add the following to 250 µL sample: 116 µL 7.5 M ammonium acetate, 4 µL glycogen, 1060 µL cold 100 % ethanol. Incubate at −40 °C for 30 min or overnight. Centrifuge at maximum speed in a benchtop centrifuge for 40 min and at 4 °C.
11. Wash pellet twice with 1 mL 70 % ethanol. Let the pellet air-dry for a minimum of 10 min and dissolve the pellet in 10 µL LoTE.
12. Determine the ditag concentration (*see* **Note 15**). Approximately 1 µg of 34-bp ditag is needed for concatenation.

3.10. Concatenation of Ditags

1. Set up a ligation reaction on ice using the gel-purified 34-bp ditags by combining 8 µL 34-bp ditag (1000–2500 ng), 1 µL 10X T4 DNA ligase buffer, 1 µL T4 ligase. Incubate for 30 min at 16 °C. Heat sample at 65 °C for 15 min to terminate the ligation reaction.
2. Add 3 µL of 6X TAE loading buffer. Centrifuge briefly the ligation mix at maximum speed and load entire sample into one well of a 10 × 8 cm × 0.75 mm 12 % TAE-polyacrylamide gel (comb with 10 lanes).
3. Conduct gel electrophoresis in 1X TAE running buffer at 90 V for 100 min. Stain the gel for 5 min in 25 mL 1X TAE containing 5 µL ethidium bromide. Visualize bands by exposure to UV light. Excise band ≥500 bp (*see* **Fig. 6**).
4. Electroelute concatemers by preparing the electroelution device in such a way that the elution chamber is 1 U-insert wide and the trap is 1 U-insert wide. Put the gel slice into the elution chamber and electroelute in 1X TAE running buffer at 100 V

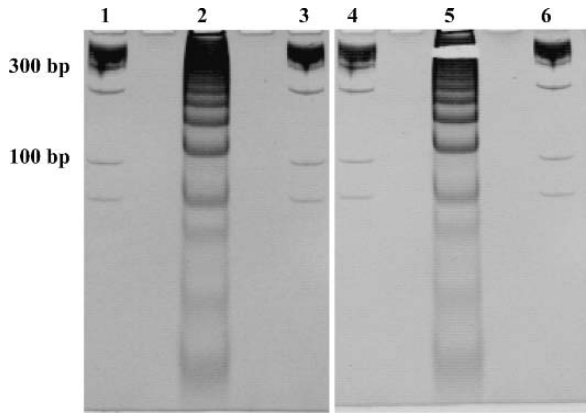


Fig. 6. Concatemers. **1,3,4,6,** GeneRuler™ 100-bp DNA ladder; **2,** concatemers; **3,** excised concatemers approx >500 bp.

and room temperature for 1 h 15 min. Remove buffer from trap (200–250 μL) and put into a new tube.
5. Extract each tube containing eluate with an equal amount of PCI. Centrifuge at maximum speed for 5 min at room temperature.
6. Transfer the upper aqueous phase (approx 200 μL) from each tube to a new tube.
7. To each 200 μL eluate add 133 μL 7.5 *M* ammonium acetate, 3 μL glycogen, and 1000 μL cold 100 % ethanol. Mix well and incubate at −40 °C for a minimum of 30 min. Centrifuge at maximum speed for 30 min at 4 °C.
8. Carefully remove and discard the supernatant. Wash pellet twice with 1 mL cold 70 % ethanol.
9. Carefully remove ethanol and air-dry pellet for 10–20 min. Resuspend pellets in a final volume of 14 μL LoTE.

3.11. Cloning of Concatemers

1. Linearize 2 μg of pZErO-1 with *Sph*I by mixing the following on ice: 2.0 μL pZErO-1 (1 μg/μL), 2.5 μL 10X Buffer 2, 1.4 μL *Sph*I (5 U/μL), 19.1 μL sterile MilliQ water. Incubate at 37 °C for 20 min. Add 175 μl LoTE to the digestion mix, and an equal volume (~200 μL) of PCI to the tube and mix vigorously. Centrifuge at maximum speed for 5 min at room temperature. Transfer the upper aqueous phase (~200 μL) to a clean tube. Add 65 μL of 7.5 *M* ammonium acetate and 600 μL 100 % ethanol. Mix well. Store tube at −40 °C for 30 min, and centrifuge at maximum speed in a benchtop centrifuge for 30 min at 4 °C.
2. Carefully remove and discard the supernatant. Wash pellet twice with 1 mL 70 % ethanol and air-dry pellet for 10 min. Resuspend pellet in 60 μL LoTE. Mix well and aliquot into six 10-μL tubes. Store the tubes at −20 °C. Analyze 2 μL of the digestion mixture and undigested pZErO-1 on a 1 % agarose gel.
3. Set up the following three ligation reaction on ice: sample (1), 14 μL concatemers, 2 μL pZErO-1/*Sph*1, 2 μL 10X T4 DNA ligase buffer, 2 μL T4 DNA ligase; no DNA (2), 1 μL pZErO-1/*Sph*1, 1 μL 10X T4 DNA ligase buffer, 2 μL T4 DNA ligase, 6 μL Sterile MilliQ water; no ligase (3), 1 μL pZErO-1/*Sph*1, 1 μL 10X T4 DNA ligase buffer, 8 μL Sterile MilliQ water.
4. Incubate for 2–3 h at 16 °C. Inactivate ligase by incubation for 30 min at 70 °C.
5. Add 10 μL sterile MilliQ water to the control reactions. Ligation reactions can be stored at −80 °C.
6. Add 80 μL LoTE to each of the ligation reactions, and 100 μL PCI and mix vigorously (*see* **Note 16**).
7. Centrifuge at maximum speed in a benchtop centrifuge for 5 min and transfer upper aqueous phase to a new tube.
8. Add 65 μL 7.5 *M* ammonium acetate, 3 μL glycogen and 500 μL 100 % ethanol. Mix well.
9. Incubate at −40 °C for 30 min and centrifuge at maximum speed in a benchtop centrifuge and 4 °C for 30 min.

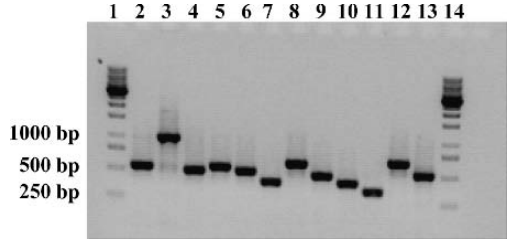


Fig. 7. Plasmid PCR products. **1,14,** GeneRuler™ 1-kb DNA ladder; **2,13,** Plasmid PCR products. PCR products larger than approx 250 bp contain concatemers.

10. Carefully discard supernatant and wash pellet four times with 1 mL 70 % ethanol.
11. Air-dry pellet for 20 min, and resuspend pellet in 20 µL LoTE.

3.12. Transformation of Escherichia coli

1. Add 1 µL of ligation reaction (sample, no DNA, or no ligase) into 50 µL of One Shot® TOP10 Electrocomp™ *E. coli* and mix gently by stirring with the pipet tip. Avoid the formation of bubbles.
2. Carefully transfer cells and DNA to a chilled 0.1-cm cuvet. Electroporate the sample at 1800 V/cm. Add 250 µL of room temperature SOC medium to the cuvet, and transfer sample to a 15-mL sterile culture tube. Incubate for 1 h at 37 °C, shaking at 200 rpm.
3. Add 750 µL of low-salt LB medium to the tube and mix. Plate 50–100 µL from each transformation on low-salt LB agar plates with Zeocin, X-gal, and IPTG, and incubate in the dark overnight at 37 °C, or at room temperature over the weekend. Store the remaining transformation reaction at 4 °C (*see* **Note 17**).
4. Pick approx 10,000 white colonies (*see* **Note 18**) for analysis into a 96-well culture plate containing 150 µL 100µg/mL LB-amp media (**Fig. 7**).

4. Notes

1. Glass pipets are recommended when working with large volumes of PCI.
2. Potato tuber contains a great deal of starch and consequently, quite a bit of tissue is used for extraction. Other, more convenient tissues require less sample.
3. 100 µg of total RNA extracted from potato tissue was used. It is imperative that undegraded RNA is used.
4. It is important that the oligo$(dT)_{25}$ beads do not dry out.
5. The oligo$(dT)_{25}$ beads might stick to the tube, so be very careful when removing the supernatant. After addition of 75 °C warm Wash Buffer C, the oligo$(dT)_{25}$ beads will no longer stick to the tube.
6. During *Mme*I digestion, the cDNA is released from the oligo$(dT)_{25}$ beads. Therefore, the supernatant contains the tags and must not be discarded.

7. First, add DNA synthesis and protein sequencing grade water to each PCR tube and then add the templates. Make a PCR reaction mix (10 reactions) and add the reagents in the following order: 10X PCR Buffer, LS DTP primer mix (175 ng/μL of each), $MgCl_2$ (25 m*M*), dNTP mix (25 m*M* of each), and Taq polymerase (5 U/μL) Mix the volume twice using a P200 pipet. Add the PCR reaction mix to each tube. Always keep PCR reaction mix and PCR tubes on ice.
8. Preheat heat block to approximately 80 °C before placing PCR reactions in heat block.
9. It is important to add the template last.
10. The centrifuge tubes are cleaned with 70 % ethanol and air dried before use.
11. The upper phase contains the ditags.
12. It may be difficult to see the pellet because it is smeared all over the back of the tube. Dissolve the pellet carefully in the final volume of 500 μL LoTE/ 200 reactions.
13. For smaller or larger samples, decrease or increase the amount proportionally.
14. According to the I-SAGE Long kit protocol, a 130-bp purification step should be conducted. This step is omitted from the protocol described here, because this step resulted in loss of valuable product. Instead of performing a 130-bp gel purification, we proceed directly to the *Nla*III digestion step, which digests 130-bp ditags to 34-bp ditags.
15. A dot blot can be used to estimate the concentration of the ditag sample. Use 1 μL ditag sample and make the following dilutions: 1:5, 1:25, and 1:125. Add 1 μL of stain (1 μL ethidium bromide + 10 mL DNA synthesis and protein sequencing grade water) to 4 μL of sample dilutions and to 4 μL of each standard solution (20 ng DNA, 10 ng DNA, 5 ng DNA, and 2.5 ng DNA). Pipet 5-μL droplets on household film, visualize under UV light, and compare samples with standards.
16. It is important to PCI-extract and ethanol-precipitate. If these steps are omitted, the cloning of concatemers into pZErO-1 will result in an insufficient number of clones.
17. The transformation reaction yields the most colonies when plated the same day of transformation. Therefore, it is recommended to plate the transformation reaction the day of transformation or at least the following day.
18. Blue–white screening helps choosing colonies with large inserts. Although white colonies with short inserts, as well as blue colonies with long inserts, are observed, all in all, the average insert length is greater for white colonies than for blue ones.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
2. Dinel, S., Bolduc, C., Boivin, A., et al. (2005) Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res*. **33(3)**, e26.

3. St. Croix, B., Rago, C., Velculescu, V. E., et al. (2000) Genes expressed in human tumor endothelium. *Science* **289**, 1197–1202.
4. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **19**, 508–512.
5. Pleasance, E. D., Marra, M. A., and Jones, S. J. M. (2003) Assesment of SAGE in transcript identification. *Genome Res.* **13**, 1203–1215.
6. Saha, S., Bardelli, A., Buckhaults, P., et al. (2001) A phosphatase associated with metastasis of colorectal cancer. *Science* **294**, 1343–1346.

2

Robust-LongSAGE (RL-SAGE)

An Improved LongSAGE Method for High-Throughput Transcriptome Analysis

Malali Gowda and Guo-Liang Wang

Summary

Serial analysis of gene expression (SAGE) is a powerful technique for large-scale transcriptome analysis in eukaryotes. However, technical difficulties in the SAGE library construction, such as low concatemer cloning efficiency, small concatemer size, and a high level of empty clones, has prohibited its widespread use as a routine technique for expression profiling in many laboratories. We recently improved the LongSAGE library construction method considerably and developed a modified version called Robust-LongSAGE, or RL-SAGE. In RL-SAGE, concatemer cloning efficiency and clone insert size were increased significantly. About 20 PCR reactions are sufficient to make a library with more than 150,000 clones. Using RL-SAGE, we have made 10 libraries of rice, maize, and the rice blast fungus *Magnaporthe grisea*.

Key Words: SAGE; LongSAGE; magnetic beads; ditags; concatemers; pZErO-1; partial digestion; *Mme*I.

1. Introduction

Serial analysis of gene expression (SAGE) was first developed by Velculescu and his colleagues at the John Hopkins University a decade ago as a new genomics tool for large-scale profiling of transcripts ***(1)***. The principle experimental steps of SAGE are isolating short tags (14–21 bp) from the 3′ ends of transcripts, generating ditags (ligating two individual tags), concatenating (ligating ditags together), and cloning these concatemers into a vector for subsequent sequencing. Specially designed computer programs are then used

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

to extract ditags and individual tags from the DNA sequences, calculate the frequency of each tag, and match the unique tags to genomic, expressed sequence tag (EST), and/or complementary DNA (cDNA) sequences. Because about 30–50 tags can be extracted from each sequence read, the SAGE method is much more cost-effective for transcript identification than EST sequencing. SAGE is also superior to microarray, because SAGE provides digital data for each transcript and is able to identify transcripts expressed at a very low level ***(2–3)***. In addition, SAGE, at least in principle, is a relatively simple procedure that can be performed in any laboratory without the use of specialized equipment. Although SAGE is a powerful tool for identifying novel transcripts and expression profiling ***(3–4)***, it has not been extensively used for gene expression studies in many organisms as a result of certain difficulties in SAGE library construction ***(5–7)***. First, almost all reports used high quantities of initial mRNA for library construction ***(1,8)*** or, when a limited amount of RNA was available, adopted PCR based pre-amplification of cDNAs and/or ditags ***(9–13)***. Second, as much as 300–1000 PCR reactions have been required for the ditag isolation, in order to obtain a sufficient amount of ditags. Third, concatemers are not easily cloned into vectors with high efficiency ***(5–6)***. Fourth, the frequent occurrence of clones without concatemers has forced researchers to adopt PCR-based colony screening of clones even though it is tedious and expensive ***(14)***. Fifth, in most of SAGE studies, the average insert size is about 300 bp, which contains approx 22 tags in the conventional SAGE method ***(6)*** or approx 15 tags in the LongSAGE method ***(7)***.

Recently, in order to successfully adapt the LongSAGE procedure, our laboratory made substantial improvements to each step in the SAGE method (Robust-LongSAGE [RL-SAGE]) ***(7)***. First, only 50 ng of initial mRNA was used to make an RL-SAGE library containing more than 150,000 clones. Second, only 20 PCR reactions were used to obtain an RL-SAGE library. Third, concatemers were partially digested with the *Nla*III enzyme before being cloned into a vector. Fourth, significant improvements in the cloning efficiency of concatemers (99 % clones with inserts) and in the insert sizes in the RL-SAGE libraries (0.8–2 kb) was obtained. Fifth, RL-SAGE clones were randomly picked directly for sequencing without colony-based PCR screening of clones. Overall, >150,000 clones (equivalent to 4.5 million tags) were obtained from each of the RL-SAGE libraries. The steps and gel images in the RL-SAGE library construction are shown in **Fig. 1**. Using these steps, we obtained three to four libraries within a month. The improved method was successfully applied to mRNA samples ***(15–17)*** from *Oryza sativa, Zea mays*, and *Magnaporthe grisea*. The RL-SAGE procedure detailed in this chapter allows a deeper transcriptome analysis in any organism, which may not be achievable using conventional

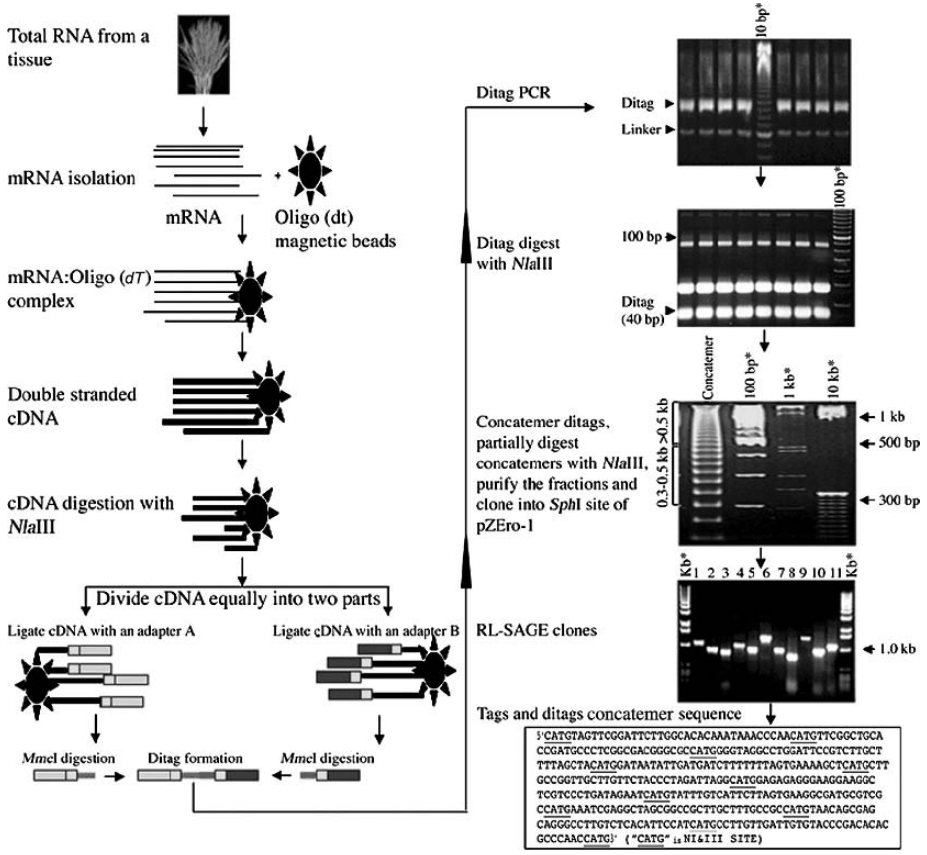


Fig. 1. Robust LongSAGE library construction procedure.

SAGE or LongSAGE procedures. It is noteworthy to mention here that a cancer genome project has sequenced >30 million tags from >298 tissues using our modifications ***(18)***.

2. Materials

2.1. Total RNA and mRNA Isolation

1. Trizol solution (Invitrogen, Carlsbad, CA).
2. Ditethylpyrocarbonate (DEPC)-treated H_2O.
3. Liquid nitrogen.
4. Chloroform.
5. Isopropanol.
6. Ethanol.
7. mRNA isolation kit (Qiagen Inc., Valencia, CA).

2.2. RL-SAGE Library Construction

1. I-SAGE/I-LongSAGE kit with magnetic stand (Invitrogen, Carlsbad, CA).
2. Siliconized (nonstick) 1.5-mL tubes (Ambion, Inc, Austin, TX).
3. Streptavidin beads (Dynal Biotech Inc., Lake Success, NY).
4. *Nla*III, *Mme*I, and *Sph*I (New England Biolabs, Inc., Beverly, MA).
5. T4 DNA ligase (5 U/μL) from USB (Cleveland, OH).
6. Polyacrylamide gel electrophoresis (PAGE)-purified oligos (Integrated DNA Technologies Inc, Coralville, IA) ***(8)***.
   Linker 1A:
   5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATATCCGACATG-3′
   Linker 1B:
   5′-TCGGATATTAAGCCTAGTTGTACTGCACCAGCAAATCC-C7 amino-modified-3′
   Linker 2A:
   5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGTCCGACATG-3′
   Linker 2B:
   5′-TCGGACGTACATCGTTAGAAGCTTGAATTCGAGCAG-C7 amino-modified-3′
7. PCR primers ***(7–8)***:
   Primer 1: 5′-biotin GTGCTCGTGGGATTTGCTGGTGCAGTACA-3′
   Primer 2: 5′-biotin GAGCTCGTGCTGCTCGAATTCAAGCTTCT-3′
8. Magnetic stand (Invitrogen, Carlsbad, CA).
9. Wash Buffer A: 10 m*M* Tris-HCl, pH 7.5 0.15 *M* LiCl, 1 m*M* ethylenediamine tetraacetic acid (EDTA), 0.1 % lithium dodecyl sulfate, 10 μg/mL glycogen
10. Wash Buffer B: 10 m*M* Tris-HCl, pH 7.5, 150 m*M* LiCl, 1 m*M* EDTA, 10 μg/mL glycogen.
11. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % sodium dodecyl sulfate (SDS), 10 μg/mL mussel glycogen.
12. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL bovine serum albumin (BSA).
13. Lysis/binding buffer: 100 m*M* Tris-HCl, pH 7.5, 500 m*M* LiCl, 10 m*M* EDTA, 1 % lithium dodecyl sulfate, 5 m*M* dithiothreitol (DTT).
14. SOC medium: 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 ***(19)***.
15. Platinum Taq DNA polymerase (Invitrogen, Carlsbad, CA).
16. Phenol:chloroform:isoamyl alcohol (25:24:1,v/v).
17. Vertical gel electrophoresis apparatus for large format gels (15 × 17 cm).
18. 40 % (w/v) acrylamide/bisacrylamide solution (29:1).
19. 10 % (w/v) ammonium persulfate.
20. *N,N,N′, N′*-tetramethylethylenediamine (TEMED).
21. 0.5-mm spacer and comb.
22. One Shot® TOP10 electrocompetent cells (Invitrogen, Carlsbad, CA).

23. Low-salt Luria-Bertani (LB) agar plates with Zeocin: 1 % Tryptone, 0.5 % yeast extract, 0.5 % NaCl, 100 μg/mL Zeocin, pH 7.5.
24. 7.5 *M* ammonium acetate.
25. Freeze medium: 2.5 % w/v LB broth, 13 m*M* KH_2PO_4, 36 m*M* K_2HPO_4, 1.7 m*M* sodium citrate, 6.8 m*M* $(NH_4)_2SO_4$, and 4.4 % v/v glycerol.
26. Zeocin antibiotic: 100 mg/mL in water.
27. TE Buffer: 10 m*M* Tris-HCl, pH 7.5, 1 m*M* EDTA.
28. pZErO-1 (1 μg/μL)(Invitrogen, Carlsbad, CA).

3. Methods

3.1. Isolation of Total RNA

1. Grind 2 g of tissue to fine powder in a mortar in liquid Nitrogen.
2. Add 15 mL of Trizol solution and incubate for 10 min.
3. Add 4 mL of chloroform incubate for 5 min and centrifuge for 20 min at 4 °C at 9000*g*.
4. Transfer supernatant into 10 mL of ice cold isopropanol, mix well, incubate on ice for 10 min and centrifuge for 15 min (9000*g*) at 4 °C.
5. Wash the RNA pellet using 15 mL of 75 % ethanol, and dry it at room temperature.
6. Dissolve total RNA in 700 μL of DEPC H_2O.

3.2. Isolation of mRNA

1. Take 700–800 μg of total RNA in an RNase-free, 1.5-mL tube, and adjust the volume to 500 μL with DECP water.
2. Add 500 μL OBB buffer preheated at 37 °C (Qiagen mRNA purification kit).
3. Add 55 μL of preheated (37 °C) Oligotex mixture.
4. Incubate the mixture at 70 °C for 3 min in a water bath, and then place it at room temperature for 10 min.
5. Centrifuge for 2 min at (9,000 g) and transfer the supernatant carefully to a new tube.
6. Add 600 μL of OW2 buffer to Oligotext:mRNA pellet.
7. Transfer the mixture to a spin column, centrifuge for 1 min at (9000*g*) and discard the flow through.
8. Wash mRNA in the column with 600 μL of OW2 buffer, centrifuge for 1 min at (9000*g*) and discard the flow through.
9. Add 100 μL of hot OEB buffer (heated at 70 °C), and centrifuge for 1 min to collect the mRNA solution.
10. Again, wash column with 100 μLof hot OEB buffer (heated at 70 °C), centrifuge for 1 min, and collect the mRNA solution.
11. Precipitate mRNA by adding 20 μL of 3 *M* sodium acetate and 1 mL of 100 % ethanol.
12. Wash the mRNA pellet with 70 % ethanol and dry the pellet.

13. Dissolve in 7 μL of RNase free DEPC water.
14. Estimate mRNA concentration by analyzing 1 μL on a 1.5 % agarose gel and by absorption at 260 nm.

3.3. Equilibration Oligo(dT) Beads in Lysis/Binding Buffer

1. Transfer 100 μL of oligo(dT) beads to a siliconized (nonstick) tube.
2. Place in a magnet for 1 min and remove the supernatant carefully with a pipet.
3. Add 500 μL lysis/binding buffer and incubate for 2 min at room temperature.
4. Place in a magnet for 1 min and remove the supernatant carefully with a pipet without disturbing beads.

3.4. Binding mRNAs to the Oligo(dT) Beads

1. Add approx 50 ng of mRNA into 1 mL of lysis buffer.
2. Add this mixture into a siliconized tube containing equilibrated oligo(dT) beads.
3. Agitate on a shaker for 30 min at room temperature.
4. Centrifuge for 30 s at (9000*g*) place in a magnet for 1 min, and remove the supernatant.
5. Wash the beads twice with 1 mL of wash buffer A.
6. Wash the beads twice with 1 mL of wash buffer B.
7. Wash the beads four times with 100 μL of 1X first strand cDNA synthesis buffer. Leave the beads suspended in the last buffer wash.

3.5. cDNA Synthesis

1. Prepare the following first stand cDNA synthesis reaction mixture (*see* **Note 1**). Eighteen microliters of 5X first strand cDNA synthesis buffer, 1 μL of RNaseOUT (40U/μL), 9 μL 0.1 *M* DTT, 4.5 μL dNTP mixture (10 μ*M* each) and 55 μL of water.
2. Place beads in a magnet for 1 min and remove the supernatant carefully and immediately add the above mixture for first strand cDNA synthesis to the beads.
3. Incubate at 37 °C for 2 min and add 3 μL of SuperScript™ II (200 U/μL) by incubating at 37 °C for 1 h (vortex beads every 15 min).
4. Then place the above first stand mixture on ice and prepare for second strand cDNA synthesis by mixing following reagents: 465 μL of water (prechilled on ice), 150 μL 5X second strand buffer, 15 μL dNTPs (10 m*M* each), 5 μL *Escherichia coli* ligase (10 U/μL, 20 μl *E. coli* DNA polymerase I (10 U/μL), and 5 μl *E. coli* RNase H (2 U/μL).
5. Mix the beads with the above mixture for second strand synthesis, and incubate at 16 °C for 2 h with intermittent mixing every 15 min.
6. Add 50 μL of 0.5 *M* EDTA (pH 7.5) and place on ice.
7. Place in a magnet for 1 min and remove the supernatant carefully, and wash beads once with 750 μL of (preheated to 75 °C) buffer C by incubating at 75 °C for 15 min.

8. Wash once more with 750 μL of buffer C, this time at room temperature (be quick, otherwise SDS precipitates).
9. Wash once with 750 μL of buffer D.
10. Wash four times with 200 μL of 1X *Nla*III buffer.
11. After the final wash, transfer the beads into a new siliconized tube and proceed with the *Nla*III digestion.

3.6. Digest cDNA With NlaIII

1. Prepare for cDNA digestion with *Nla*III by mixing the following reagents. 172 μL water, 2 μL BSA, 20 μL 10X New England Biolabs buffer 4 and 6 μL *Nla*III (1 U/μL).
2. Incubated the above reaction with the beads at 37 °C for 2 h while mixing (flicking or vortexing) at 15-min intervals. After the incubation, place the tubes on ice and prepare for washing the beads.
3. Place the tube in a magnet for 1 min and transfer the supernatant carefully to a new tube. Precipitate the supernatant overnight at −80 °C by adding, to 300 μL, 3 μL glycogen, 133 μL 7.5 *M* ammonium acetate, and 1 mL of 100 ethanol. Check cDNA synthesis and *Nla*III digestion by analyzing 5 μL of the reaction on a 2 % agarose gel. A smear ranging from 100 to 1000 bp should be visible. This indicates successful cDNA synthesis and *Nla*III digestion.
4. Wash the beads twice with 750 μL Wash Buffer C (prewarmed at 37 °C).
5. Wash four times with 750 μL of Wash Buffer D.
6. Wash twice with 150 μL of 1 X ligase buffer.
7. Add 200 μL of 1X ligase buffer to beads.
8. Then equally divide beads in two parts (100 μL each) and label as tube A and tube B.
9. Place in a magnet for 1 min and remove the supernatant. Add 1X ligase buffer.

3.7. Ligation of Adapters A and B to the cDNAs Fragments

1. Mix the following reagents: 1.5 μL Adapter A, 2 μL of 10X ligase buffer and 14 μL water in one tube (A reaction mix) and 1.5 μL Adapter B, 2 μL of 10X ligase buffer and 14 μL water in another tube (B reaction mix). Place tube A or B in a magnet and remove the supernatant. Resuspend the beads in A or B reaction mix, respectively.
2. Mix the beads well, heat both tubes to 50 °C for 2 min, then cool them to room temperature for 15 min before placing these tubes on ice.
3. Add 2.5 μL of T_4 DNA ligase enzyme (5 U/μL, USB) to each tube, mix well, and incubate at 16 °C for 3 h while flicking the tubes every 15 min. Allow the ligation process to continue overnight.
4. Place tubes on a magnetic stand for 2 min and remove the solution. (**Note:** do not discard the ligation solution. Instead, load entire sample on a 4 % agarose gel

alongside equal amounts of Adapter 1 and 2. In the ligation supernatant, only a faint, unligated band (surplus adapters) compared to adapter bands should be apparent.
5. Wash beads four times with 500 µL of buffer D.
6. Resuspend beads in 200 µl 1X *Mme*I buffer with 1X *S*-adenosyl-methionine (SAM).

3.8. cDNA Digestion With MmeI Enzyme

1. Mix in two tubes the following reagents, 30 µL 10X NEB buffer 4, 3 µL 100X SAM, 20 µL *Mme*I (2 U/µL), and 220 µL water. Place the beads in a magnet for 1 min and remove the supernatant. Add the *Mme*I digestion mix to each tube.
2. Mix well and incubate at 37 °C for 3 h.
3. Place the two reactions in a magnet for 1 min and remove the supernatant (containing tags) carefully from tubes A and B, and pool them together. Discard the beads.
4. Extract the tag solution with an equal volume of phenol:chloroform:isoamyl alcohol (24:24:1), mix, and centrifuge for 10 min at (9000*g*) and transfer the upper aqueous phase to a new tube.
5. Precipitate DNA by adding (for 300 µL solution) 133 µL of 7.5 *M* ammonium acetate, 3 µL mussel glycogen and 1 mL of 100 % ethanol) and incubate overnight at −80 °C.
6. Centrifuge for 60 min at 4 °C and remove supernatant.
7. Wash the pellet twice with 1ml of 75 % ethanol.
8. Dry the DNA pellet for 5–10 min in a fume hood with the lid open (do not dry completely).
9. Dissolve the DNA pellet in 5 µL of sterile H_2O for 10 min at room temperature.

3.9. Ditag Formation

1. To the 5 µL of digested tag solution, add 1 µL 10X ligase buffer, 1.25 µL T4 DNA ligase, and 2.75 µL water.
2. Incubate at 16 °C overnight.

3.10. PCR Amplification of Ditags

1. Dilute the ditag DNA (ligated product) with sterilized water (1:100).
2. Mix the following reagents and add to 1 µL of the diluted ditags: 34 µL water, 5 µL 10X PCR buffer, 5 µL dNTP mix (2.5 m*M* each), 1.5 µL $MgCl_2$ (50 m*M*), 2 µL Bio-forward primer (350 ng/µl), 2 µL Bio-reverse primer (350 ng/µL), and 0.5 µL Platinum Taq DNA polymerase (5 U/µL).
3. Mix the PCR reactions well and perform the following PCR cycling profile. After initial denaturation at 94 °C for 2 min, 27 cycles of 94 °C for 30 s, 55 °C for 1 min, and 70 °C for 1 min should be followed by 5 min at 72 °C.
4. Perform 20 PCR reactions and check the ditag band intensity on a 4 % agarose gel by loading 5 µL of PCR product along with 100-bp ladder (if bands are faint, increase to up to 50 PCR reactions per library). A bright, approx 140-bp ditag band

and a faint, 100-bp linker band should be observed. Pooling and precipitating DNA from multiple PCR reactions dramatically reduces sharpness of PCR bands (*see* **Note 2**).

3.11. PAGE Purification of PCR Amplicons of Ditag Band

1. Prepare a 12 % PAGE gel using a medium-sized electrophoresis gel unit (15 × 17 cm).
2. Prerun PAGE at 100 V for 30 min.
3. Wash wells with buffer, load one PCR sample into each lane, and carry out the electrophoresis at 75 V for 12 h until xylene cyanol dye reaches three-fourths of the gel.
4. Carefully remove the PAGE gel and stain with ethidium bromide (1.5 μg/mL) in 200 mL buffer solution for 25 min.
5. Quickly visualize bands by short ultraviolet exposure and excise the ditag band (~140 bp).
6. Transfer gel slice into a 0.5-mL tube (bottom punctured by 18-G needle) and place this tube in a 1.5-mL tube.
7. Centrifuge for 5–10 min at (9000*g*) until gel slice transfers from the 0.5-mL tubes to the 1.5-mL tubes.
8. Add 250 μL TE buffer and 50 μL of 7.5 *M* ammonium acetate.
9. Vortex the mixture and incubate at 65 °C for 2 h. Alternatively, tubes can remain at 4 °C overnight, and be vortexed and incubated at 65 °C for 30 min the following day.
10. Transfer gel paste into the SNAP column.
11. Centrifuge for 5 min at (9000*g*).
12. Extract the flow-through with an equal volume of phenol:chloroform:isoamyl alcohol) (24:24:1), centrifuge, and transfer the upper aqueous phase to a new tube.
13. To 300 μL of DNA solution, add 133 μL of 7.5 *M* ammonium acetate, 3 μL mussel glycogen, and 1 mL of 100 % ethanol.
14. Incubate overnight at −80 °C.
15. Centrifuge for 60 min at 4 °C and wash the pellet twice with 1 mL of 75 % ethanol.
16. Dry the DNA pellet for 5–10 min at room temperature (do not dry completely) and dissolve the DNA in 60 μL of H_2O.

3.12. Digest Ditags With NlaIII

1. Prepare the following reaction mixture and add to the above ditag DNA 186 μL water, 30 μL 10X NEB buffer 4, 4 μL 100X BSA, and 20 μL *Nla*III (20 U/μL).
2. Mix well and divide the above mixture into three tubes (100 μL each) and incubate at 37 °C for 3 h.
3. The ditag DNA digestion may be checked by loading 5 μL of digested sample on a 4 % agarose gel. If digestion was successful, a 40-bp ditag band, a 50-bp linker

band, a faint, 90-bp band from partially digested ditags, and a faint, 140-bp band from undigested ditags can be seen on a 4 % agarose gel.

3.13. PAGE Purification of NlaIII-Digested Ditags

1. Prepare a 16 % PAGE gel, load the entire digested sample in six lanes, and resolve at 75 V for 10 h until the big dye reaches to three-fourths from the top (42 bp move along with the big dye).
2. Remove gel and stain with ethidium bromide as described in **Subheading 3.11., step 4** above.
3. Isolate the ditag band (~40 bp) from the gel and incubate gel slice at 37 °C for 2 h or incubate at 4 °C overnight (do not incubate ditags at 65 °C; this would denature them).
4. Precipitate the ditags as in **Subheading 3.11., steps 12** and **13**.
5. Dissolve ditag DNA in 100 μL of TE.

3.14. Purify Ditags Using Streptavidin Beads

1. Transfer 100–400 μL of Dynal streptavidin beads into a 1.5-mL siliconized tube.
2. Remove supernatant from beads using magnetic stand, as described in **Subheading 3.7., step 4** above.
3. Add ditag DNA (100 μL) to Dynal Streptavidin beads, and shake for 30 min at room temperature.
4. Place on a magnet for 1 min and transfer the ditag-containing supernatant to another tube. Extract with an equal volume of PCI (refer to **Subheading 3.11., step 12**), centrifuge, and transfer the supernatant to a new tube.
5. Precipitate ditag DNA as described in **Subheading 3.11., step 13**.
6. Dissolve ditag DNA in 5 μL H_2O at room temperature for 15–20 min.

3.15. Formation of Concatemers

1. Set up following mixture for the ditag ligation reaction and incubate at 16 °C for 3 h: 5 μL ditag DNA, 1 μL 10X ligase buffer, and 1.25 μL T4 DNA ligase (5 U/μL).
2. Check concatemer formation by resolving 0.5 μL of the ligation mixture on a 2 % agarose gel.
3. If the ligation process was successful, a smear should be seen from 100–500 bp.

3.16. Partial Digestion of Concatemer DNA With NlaIII

1. To the ligation reaction, add the following reaction mixture: 7 μL water, 2 μL 10X *Nla*III buffer, and 0.3 μL 100X BSA and incubate this mixture at 37 °C for 5 min.
2. Add 1 μL of *Nla*III (10 U/μL), mix well and quickly by flicking the tube, and incubate for just 1 min at 37 °C. Do not allow more than 1 min of digestion (*see* **Note 3**).
3. Inactivate *Nla*III at 75 °C for 15 min and then place the tube on ice.

3.17. Concatemer DNA Purification on a 6 % PAGE

1. Prepare a 6 % PAGE gel and prerun for 30 min..
2. Load the entire *Nla*III-digested ligation mixture in a single lane and load the other lanes with 1 kb and 100 bp marker DNA.
3. Resolve concatemers on a 6 % PAGE gel at 75 V for 12 h until the big dye reaches to three-quarters from the top.
4. Purify the concatemer smear above 500 bp as above (*see* **Subheading 3.11., step 13**).
5. Incubate the gel slice at 65 °C for 2 h, complete SNAP-column purification, and precipitate by following above steps (*see* **Subheading 3.11., step 13**).
6. Add 10 μL of H_2O to concatemer DNA pellet.

3.18. Preparation of Vector DNA for Ligation

1. Digest about 2 μg of pZErO-1 vector DNA with *Sph*I by combining the following reagents: 2 μL pZErO-1, 2.5 μL 10X NEB buffer 2, 19 μL water, and 1.5 μL *Sph*I (5 U/μL).
2. Incubate at 37 °C for 25 min.
3. To inactivate *Sph*1, heat the mixture to 70 °C for 20 min, extract with an equal volume of PCI as in **Subheading 3.11., step 12**, centrifuge, and remove the supernatant.
4. Precipitate plasmid DNA as in **Subheading 3.11., step 13**.
5. Dissolve pZErO-1 DNA in 60 μL of H_2O.

3.19. Ligation of Concatemer

Clone concatemers by combining the following: 1.5 μL water, 1.5 μL concatemer DNA from **Subheading 3.17.**, 1 μL *Nla*I-digested pZErO-1, 0.5 μL 10X ligase buffer, and 0.5 μL T4 DNA ligase (5U/μL) and incubate the reaction at 16 °C overnight.

3.20. Transformation

1. Mix about 0.5 μL of ligation mixture with 20 μL of competent cells (e.g., One Shot® TOP10 Electrocomp™ *E. coli*).
2. Transform bacteria by electroporation.
3. Transfer transformation mixture into 1 mL of SOC medium and incubate at 37 °C for 45 min.
4. Plate 100 μL of mixture on low-salt LB (50 μg/mL Zeocin), and incubate overnight at 37 °C.
5. Randomly pick 20 concatemer clones and check the average insert size by colony PCR.
6. If insert size is satisfactory, randomly pick about 5000 clones for sequencing that may yield approx 150–200,000 individual tags (*see* **Note 13**, **14** and **15**).

Acknowledgments

We thank Chatchawan Jantasuriyarat for his help during RL-SAGE library construction. We also thank to Rose Palumbo for the critical reading of this chapter. The methods described here were developed with supports from a National Science Foundation-Plant Genome Research Program (# 0115642).

4. Notes

1. Each step in the RL-SAGE procedure is technically challenging. The steps from cDNA synthesis to concatemer purification are critical. Therefore, it is recommended that users follow the procedure continuously until the purification of concatemers from a PAGE gel.
2. Partial digestion of concatemers will yield better results. This step has significantly reduced the number of PCRs required for ditag amplifications from 300 to 20.
3. If concatemers were digested partially with the *Nla*III enzyme, we found a dramatic improvement in insert sizes (from 800 to 2000 bp).
4. Almost all clones were found to have an insert size of around 1.0 kb. These improvements overcame the need for PCR to screen concatemers for sequencing.
5. Compared to 300-bp inserts (concatemer clones) in conventional SAGE and LongSAGE libraries (22 tags in conventional SAGE and 14 tags in LongSAGE), we obtained 1.0 kb inserts (71 tags in conventional SAGE or 47 tags in LongSAGE).
6. Most reports characterized short concatemers (approx 300 bp or 22 tags), which is not a practical way to achieve quantitative deep sampling of transcript tags when following other procedures.
7. Most reports followed colony PCR-based screening of concatemers for large-scale sequencing because of high rates of empty and smaller concatemers.
8. More than 150,000 clones (concatemers) were obtained in each RL-SAGE library, which is equivalent to about 4.5 million transcript tags (21 bp).
9. Only 5000 clones from each library were sequenced as a result of the high cost of sequencing, and from these, about 150–200,000 individual tags were found.
10. A library construction takes about 2 wk if the procedure is followed strictly.
11. Using this method, more than 2 million tags (21 bp) from 10 SAGE libraries were generated within a single year, as compared to 7 million tags from hundreds of SAGE libraries since the first report of SAGE in 1995.
12. SAGE methodology can be used easily as a common laboratory tool for transcriptome and gene discovery by following our modifications as we report in this procedure.
13. Based on a literature survey, only a few papers have been published using the LongSAGE procedure, but more than 200 SAGE library papers have been published in model organisms (www.sagenet.org). This indicates that the SAGE method has not been fully adopted by the plant research community as compared to the medical community.

14. We use SAGEspy software for tag extraction and analysis ***(16)***.
15. Recently, RL-SAGE modifications have been adopted in medical research community by sequencing >30 million tags from >298 tissues ***(18)***.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270,** 484–487.
2. Boon, K., Osorio, E. C., Greenhut, S. F., et al. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA* **99,** 11,287–11,292.
3. Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J. D., and Wang, S. M. (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci. USA* **99,** 12,257–12,262.
4. Sun, M., Zhou, G., Lee, S., Chen, J., Shi, R. Z., and Wang, S. M. (2004) SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics* **5,** 1.
5. Powell, J. (1998) Enhanced concatemer cloning: a modification to the SAGE (serial analysis of gene expression) technique. *Nucleic Acids Res.* **26,** 3445–3446.
6. Kenzelmann, M. and Muhlemann, K. (1999) Substantially enhanced cloning efficiency of SAGE (serial analysis of gene expression) by adding a heating step to the original protocol. *Nucleic Acids Res.* **27,** 917–918.
7. Gowda, M., Jantasuriyarat, C., Dean, R., and Wang, G. L. (2004) Robust-LongSAGE (RL-SAGE) for both gene discovery and transcriptome analysis. *Plant Physiol.* **134,** 890–897.
8. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **20,** 508–512.
9. Peters, D. G., Kassam, A. B., Yonas, H., O'Hare, E. H., Ferrell, R. E., and Brufsky, A. M. (1999) Comprehensive transcript analysis in small quantities of mRNA by SAGE-Lite. *Nucleic Acids Res.* **27,** e39.
10. Datson, N. A., van der Perk-de Jong, J., van den Berg, M. P., de Kloet, E. R., and Vreugdenhil, E. (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res.* **27**, 1300–1307.
11. Virlon, B., Cheval, L., Buhler, J. M., Billon, E., Doucet, A., and Elalouf, J. M. (1999) Serial microanalysis of renal transcriptomes. *Proc Natl. Acad. Sci. USA* **96,** 15,286–15,291.
12. Neilson, L., Andalibi, A., Kang, D., et al. (2000) Molecular phenotype of the human oocyte by PCR-SAGE. *Genomics* **63,** 13–24.
13. Vilain, C., Libert, F., Venet, D., Costagliola, S., and Vassart, G. R. (2003) Small amplified RNA-SAGE: an alternative approach to study transcriptome from limiting amount of mRNA. *Nucleic Acids Res.* **31,** e24.
14. Fujii, S. and Amrein, H. (2002) Genes expressed in the Drosophila head reveal a role for fat cells in sex-specific physiology. *EMBO J*. **21,** 5353–5363.

15. Gowda, M., Venu, R. C., Raghupathy, M. B., et al. (2006) Deep and Comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE and oligoarray methods. *BMC Genomics* **7**, 310.
16. Gowda, M., Venu, R. C., Jia, Y., et al. (2007) Use of robust-long serial analysis of gene expression to identify novel fungal and plant genes involved in host-pathogen interactions. In *Plant-Pathogen Interactions* (Ronald, P. C., ed.). Humana press, Totowa, NJ, *Methods Mol. Biol.* **354**, 31–44.
17. Gowda, M., Venu, R. C., Li, H., et al. (2007) *Magnaporthe grisea* Infection Triggers RNA Variation and Antisense Transcript Expression in Rice. *Plant Physiol.* (in press).
18. Khattra, J., Delaney, A. D., Zhao, Y., et al. (2007) Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Res.* **17**, 108–116.
19. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual, Vols. 1–3.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

3

aRNA-LongSAGE

SAGE With Antisense RNA

Anna M. Heidenblut

Summary

In order to generate serial analysis of gene expression (SAGE) libraries from very small samples such as microdissected cells, the starting material must first be amplified via PCR or linear amplification of RNA. In microarray experiments, it has been shown that linear amplification of RNA can be used to generate reliable gene expression profiles and leads to the detection of expression differences that are not seen with nonamplified starting material. As the product of the amplification is amplified antisense RNA (aRNA), linear amplification of RNA cannot be used in combination with the conventional SAGE protocol. The aRNA-LongSAGE protocol described herein is an adaptation of the MicroSAGE protocol to the use of aRNA as starting material.

Key Words: RNA amplification; T7 RNA polymerase; amplified antisense RNA; aRNA; aRNA-LongSAGE; expression profiles.

1. Introduction

Amplified antisense RNA-Long serial analysis of gene expression (aRNA-LongSAGE) is a modification of the conventional SAGE protocol that allows the generation of SAGE libraries from very small sample sizes such as microdissected cells ***(1)***. As little as 40 ng of total RNA are sufficient to generate an aRNA-LongSAGE library. This is achieved by linear amplification of RNA, which is carried out prior to the synthesis of the SAGE library.

Linear amplification of RNA is a method routinely used in gene expression profiling via microarrays ***(2)***. It starts with a cDNA synthesis using a modified

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

oligo(dT) primer that adds the T7 RNA polymerase promoter to the 3′ end of the cDNA. In vitro transcription of this cDNA with T7 RNA polymerase yields aRNA. This technique introduces less amplification bias than PCR-based complementary DNA (cDNA) amplification protocols ***(3)***. Furthermore, the use of aRNA in differential gene expression analysis leads to the detection of expression differences that are not observed when using nonamplified RNA as starting material ***(4,5)***. The majority of these additional expression differences can be verified by quantitative real-time PCR ***(4)***. The aRNA obtained by linear amplification of RNA cannot be used in combination with the standard SAGE protocol, as the latter needs sense RNA for the cDNA synthesis, which is the first step of library generation.

The aRNA-LongSAGE protocol described herein uses a modified cDNA synthesis to adapt the SAGE procedure to the use of antisense RNA. This is done by using a random primer for the cDNA first strand synthesis (*see* **Fig. 1**). This so called "SAGErandom" primer consists of six random nucleotides and

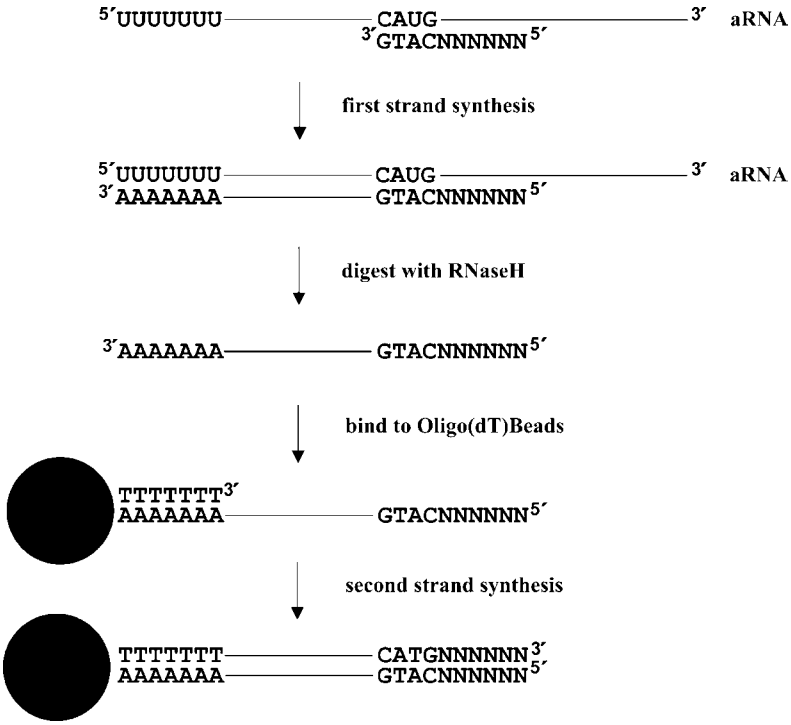


Fig. 1. Schematic overview of the cDNA synthesis in the amplified antisense RNA-Long serial analysis of gene expression protocol.

the recognition site of the SAGE anchoring enzyme, *Nla*III. The *Nla*III site was included to specifically reverse transcribe only those RNA molecules that are accessible to the SAGE procedure. After first strand synthesis, the RNA is digested with RNaseH and the resulting cDNA first strand can be hybridized to oligo(dT) beads due to its polyA tail. The oligo(dT) sequence on the beads serves as a primer for the synthesis of the second cDNA strand. After second strand synthesis, all further steps are done according to the MicroSAGE protocol ***(6)*** adapted for LongSAGE ***(7)***. Some minor modifications were introduced to improve the yield of ditags and the length of concatemers (*see* **Fig. 2** for gel photographs of an aRNA-LongSAGE library).

Using the LongSAGE rather than the conventional SAGE protocol improves the annotation of SAGE tags, as LongSAGE tags are 21-bp long whereas conventional SAGE tags are only 14-bp long. However, the modified cDNA synthesis used in the aRNA-LongSAGE protocol can be combined with the conventional SAGE protocol as well as with the LongSAGE protocol.

2. Material

2.1. Preparation of Amplified Antisense RNA

RNA amplification kit, e.g. MessageAmp from Ambion (Huntingdon, UK).

2.2. cDNA Synthesis

1. Diethylpyrocarbonate (DEPC)-treated water: add 2 mL DEPC to 1 L of water (*see* **Note 1**), shake for 30 min, and autoclave.
2. SAGErandom oligonucleotide 5′-NNN NNN CATG-3′, 125 ng/μL.
3. dNTP mix, 10 m*M* each, store aliquots at −20 °C.
4. Dry ice and wet ice.
5. 5X First Strand Buffer, 0.1 *M* dithiothreitol (DTT), RNaseOUT, and Super Script III Reverse Transcriptase (all from Invitrogen, Karlsruhe, Germany).
6. 5X Second Strand Buffer: 94 m*M* Tris-HCl, pH 6.9, 453 m*M* KCl, 23 m*M* $MgCl_2$, 50 m*M* $(NH_4)_2SO_4$, 0.75 m*M* β-NAD.
7. RNaseH (5 U/μL; USB, Cleveland, OH) is diluted with Second Strand Buffer to a final concentration of 2 U/μL.
8. 1.5-mL sterile, siliconized microcentrifuge tubes (Ambion).
9. Oligo$(dT)_{25}$Beads (Dynal Biotech Hamburg, Germany) and a magnetic stand.
10. Overhead shaker.
11. Binding Buffer and Washing Buffer B from Dynal (Dynabeads®mRNA Purification Kit).
12. *Escherichia coli* DNA Polymerase I (11.8 U/μL; USB) *E. coli* DNA Ligase (10 U/μL; USB) and T4 DNA Polymerase (3 U/μL; New England Biolabs [NEB], Frankfurt a.M., Germany).
13. Thermomixer and thermocycler for incubation steps.

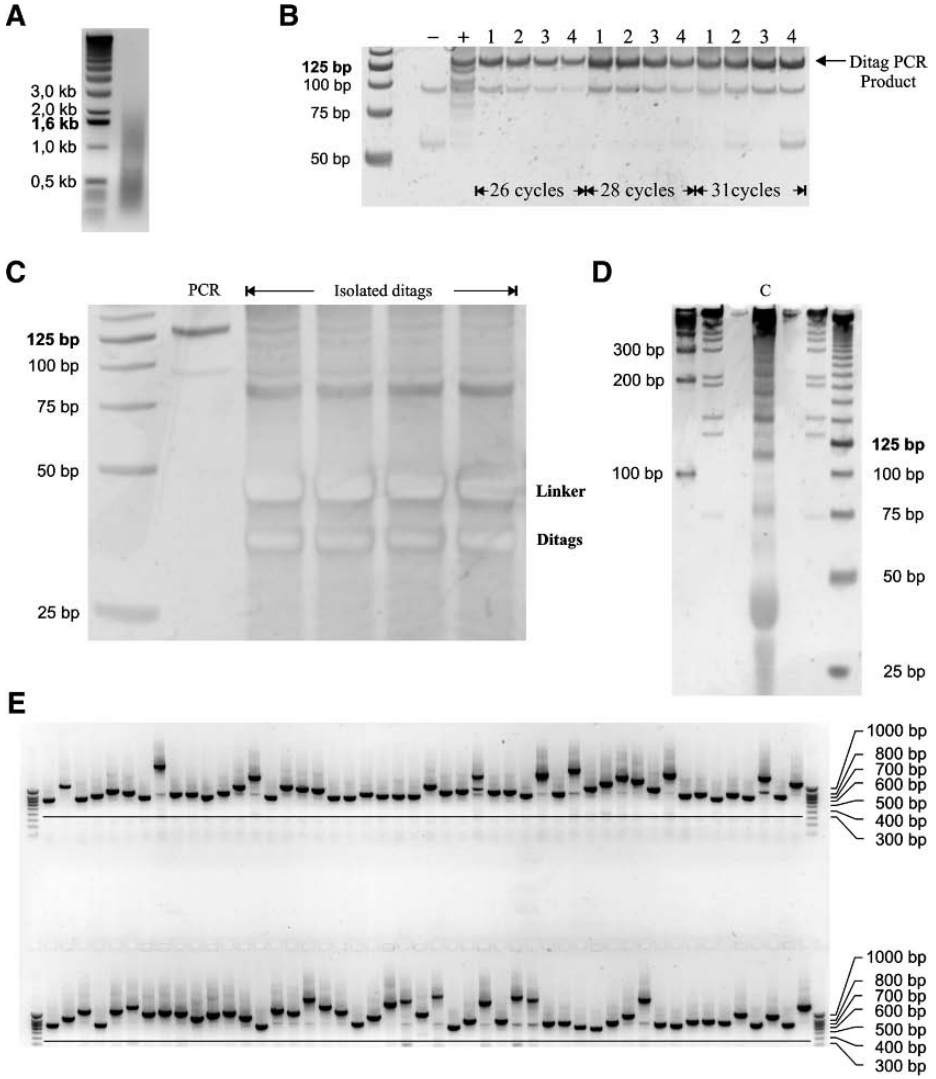


Fig. 2. Representative gels of an amplified antisense RNA-Long serial analysis of gene expression library construction. **A,** 0.5-μL first strand cDNA on a 1 % agarose gel; **B,** ditag PCR on a 12 % polyacrylamide gel (−, negative control; +, positive control); **C,** isolated ditags on a 12 % polyacrylamide gel (PCR, 5 μL of ditag PCR prior to Hsp92II digestion, 4 of 7 lanes with isolated ditags are shown in the photograph); **D,** concatemers (C) on a 8 % polyacrylamide gel; **E,** insert PCR on a 1.5 % agarose gel, the horizontal line shows the product size that corresponds to an empty cloning vector.

14. 0.5 *M* ethylenediamine tetraacetic acid (EDTA).
15. Buffer BW: 1 *M* NaCl, 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA.
16. BW/BSA: Buffer BW containing 0.1 mg/mL bovine serum albumin (BSA) (NEB).
17. 10X Buffer K from Promega (Ingelheim, Germany).

2.3. Cleavage of cDNA With the Anchoring Enzyme (Hsp92 II)

1. 100X BSA (10 mg/mL; NEB).
2. Hsp92II (10 U/μL; *see* **Note 2**) and 10X Buffer K (both from Promega).
3. 5X Ligase Buffer from Invitrogen.

2.4. Ligating Linkers to Bound cDNA

1. Linker1A:
   5′-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA TCC GAC ATG-3′
   Linker1B:
   5′-TCG GAT ATT AAG CCT AGT TGT ACT GCA CCA GCA AAT CC(Amino C7)-3′
   Linker2A:
   5′-TTT CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC GTC CGA CAT G-3′
   Linker2B:
   5′-TCG GAC GTA CAT CGT TAG AAG CTT GAA TTC GAG CAG(Amino C7)-3′
   Linker oligonucleotides are dissolved in LoTE to a final concentration of 350 ng/μL and stored at −20 °C.
2. LoTE: 3.0 m*M* Tris-HCl, pH7.5, 0.3 m*M* EDTA.
3. 10X polynucleotide kinase buffer, 10 m*M* ATP and T4 polynucleotide kinase (10 U/μL; all from NEB)
4. HC T4 Ligase (5 U/μL) and 5X Ligase Buffer (both from Invitrogen)
5. 10X Buffer 4 from NEB

2.5. Release of cDNA Tags Using the Tagging Enzyme MmeI

1. 32 m*M* *S*-adenosylmethionine (SAM; NEB).
2. *Mme*I (2 U/μL) and 10X Buffer 4 from NEB.
3. PC8: Roti®-Phenol/Chloroform/Isoamyl alcohol, pH 7.5–8.0 from Roth (Karlsruhe, Germany), store at 4 °C.
4. 10 *M* ammonium acetate solution.
5. Glycogen (20 mg/mL) (store at −20 °C).
6. Ethanol.

2.6. Ligating Tags to Form Ditags

HC T4 Ligase (5 U/μL) and 5X Ligase Buffer (both from Invitrogen).

2.7. PCR Amplification of Ditags

1. 10X BV-Mg Buffer: 670 m*M* Tris-HCl, pH 8.8, 167 m*M* $(NH_4)_2SO_4$, 67 m*M* $MgCl_2$, 100 m*M* β-mercaptoethanol.
2. Dimethylsulfoxide (DMSO), store at −20 °C.
3. dNTP mix, 10 m*M* each, store aliquots at −20 °C.
4. Oligonucleotides:
   Primer 1: 5′-GTG CTC GTG GGA TTT GCT GGT GCA GTA CA-3′
   Primer 2: 5′-GAG CTC GTG CTG CTC GAA TTC AAG CTT CT-3′.
   PCR primers are resuspended in LoTE to a final concentration of 350 ng/μL and stored at −20 °C.
5. Taq polymerase.
6. 40 % acrylamide/bis solution (19:1, this is a neurotoxin when unpolymerized; handle with care) and *N,N,N′,N′*-tetramethylethylendiamine (TEMED).
7. Ammonium persulfate: prepare 10 % solution in water and store at 4 °C.
8. Molecular weight marker for gel electrophoresis: 25-bp ladder from Invitrogen, diluted to a final concentration of 50 ng/μL.
9. Running buffer TAE: 40 m*M* Tris, 20 m*M* acetic acid, 2 m*M* EDTA, pH 7.5.
10. Loading buffer: 20.0 % (w/v) Ficoll® 70, 1.6 % (v/v) glycerol, 0.01 % (w/v) laurylsarcosine, 0.001 % (w/v) xylencyanole, 0.001 % (w/v) bromophenol blue in TAE.
11. Staining solution: dilute 5 μL SYBR Green 1 concentrate (BioWhittaker, Rockland, ME, USA; SYBR Green 1 is toxic, handle with care) with 50 mL TAE. Prepare staining solution fresh as required; SYBR Green I is not stable in aqueous solution.
12. 15-mL polypropylene tubes and 5-mL glass pipet for PC8 extraction of amplified ditags.
13. PC8: Roti®-Phenol/Chloroform/Isoamyl alcohol, pH 7.5–8.0 from Roth, store at 4 °C.
14. 15-mL centrifugation tubes.
15. 10 *M* ammonium acetate solution.
16. Glycogen (20 mg/mL) (store at −20 °C).
17. Ethanol.

2.8. Isolation of Ditags

1. 100X BSA, 10 mg/mL.
2. Hsp92II (10 U/μL) and 10X Buffer K (both from Promega).
3. PC8: Roti®-Phenol/Chloroform/Isoamyl alcohol, pH 7.5–8.0 from Roth, store at 4 °C.
4. 10 *M* ammonium acetate solution, glycogen (20 mg/mL, store at −20 °C) and ethanol.
5. TE solution (Invitrogen).
6. 40 % acrylamide/bis solution (19:1; this is a neurotoxin when unpolymerized; handle with care) and TEMED.

7. Ammonium persulfate: prepare 10 % solution in water and store at 4 °C.
8. Glycerol.
9. Molecular weight marker for gel electrophoresis: 25-bp ladder from Invitrogen, diluted to a final concentration of 50 ng/μL.
10. Running buffer TAE: 40 m*M* Tris, 20 m*M* acetic acid, 2 m*M* EDTA, pH 7.5.
11. Staining solution: dilute 5 μL SYBR Green 1 concentrate (BioWhittaker, SYBR Green 1 is toxic, handle with care) with 50 mL TAE. Prepare staining solution fresh as required; SYBR Green I is not stable in aqueous solution.
12. Electroelution device (Elutrap system from Schleicher & Schüll, Dassel, Germany).
13. Chloroform.
14. 3 *M* sodium acetate solution, pH 5.2.
15. Glycogen (20 mg/mL) (store at −20 °C)
16. Ethanol.

2.9. Concatenation of Ditags

1. HC T4 Ligase (5 U/μL) and 5X Ligase Buffer (both from Invitrogen).
2. PC8: Roti®-Phenol/Chloroform/Isoamyl alcohol, pH 7.5–8.0 from Roth, store at 4 °C.
3. 10 *M* ammonium acetate solution
4. Glycogen (20 mg/mL) (store at −20 °C)
5. Ethanol.
6. 40 % acrylamide/bis solution (19:1; this is a neurotoxin when unpolymerized; handle with care) and TEMED.
7. Ammonium persulfate: prepare 10 % solution in water and store at 4 °C.
8. Loading buffer: 20.0 % (w/v) Ficoll 70, 1.6 % (v/v) glycerol, 0.01 % (w/v) laurylsarcosine, 0.001 % (w/v) xylencyanole, 0.001 % (w/v) bromphenol blue in TAE.
9. Molecular weight marker: Smart Ladder Short Fragments from Eurogentec, Seraing, Belgium.
10. Running buffer TAE: 40 m*M* Tris, 20 m*M* acetic acid, 2 m*M* EDTA.
11. Staining solution: dilute 5 μL SYBR Green 1 concentrate (BioWhittaker, SYBR Green 1 is toxic, handle with care) with 50 mL TAE. Prepare staining solution fresh as required; SYBR Green I is not stable in aqueous solution.
12. Electroelution device (Elutrap system from Schleicher & Schüll).
13. Chloroform.
14. 3 *M* sodium acetate solution, pH 5.2.

2.10. Cloning Concatemers

1. pZErO-1 Supercoiled (1 μg/μL; part of the Zero Background Cloning kit from Invitrogen).
2. 10X Buffer 2 (NEB) and *Sph*I (5 U/μL; NEB)

3. LoTE (s.2.4.) and TE (Invitrogen)
4. PC8: Roti®-Phenol/Chloroform/Isoamyl alcohol, pH 7.5–8.0 from Roth, store at 4 °C.
5. 10 *M* ammonium acetate solution.
6. Glycogen (20 mg/mL) (store at −20 °C).
7. Ethanol.
8. 5X Ligase Buffer (Invitrogen).
9. T4 DNA Ligase (1 U/μL; Invitrogen).

2.11. Transformation of Bacteria

1. Electrocompetent *E. coli* Top10 bacteria (part of the Zero Background Cloning kit from Invitrogen).
2. Electroporation device.
3. Luria-Bertani (LB) broth (Sigma, Taufkirchen, Germany).
4. Zeocin (100 mg/mL; Invitrogen) (store at −20 °C, this antibiotic is light-sensitive).
5. 5-Brom-4-chlor-3-indoxyl-β-D-galactoside (X-Gal) (Roth, store at −20 °C).
6. LB-Zeocin-X-Gal-Agar: 50 μg/mL Zeocin, 80 μg/mL X-Gal, 1.5 % (w/v) agar in LB broth.

2.12. Insert-PCR

1. 5X RDA buffer: 335 m*M* Tris-HCl, pH 8.8, 80 m*M* $(NH_4)_2SO_4$, 50 m*M* β-mercaptoethanol, 0.5 mg/mL BSA.
2. 50 m*M* magnesium chloride solution.
3. Oligonucleotides:
   Insert_for: 5′-CTG GTT AAC CTT ACT GGC TGA GTT AGC TCA CTC ATT AGG CAC-3′
   Insert_rev: 5′-TGT AAA ACG ACG GCC AGT TAC GAC TCA CTA TAG GGC GAA TTG-3′.
4. 10 m*M* dNTP-Mix (10 mM each; Promega).
5. Taq polymerase.

3. Methods

3.1. Preparation of Amplified Antisense RNA

Several companies sell kits for RNA amplification via T7 promoter-driven *in vitro* transcription of cDNA. All protocols start with cDNA synthesis using a modified oligo(dT) primer containing the T7 RNA polymerase promoter. After purification of the cDNA, the *in vitro* transcription is carried out followed by purification of the aRNA. The incubation time for the *in vitro* transcription varies between protocols. Longer incubation times give a higher yield of aRNA but might also lead to degradation of part of the aRNA. For the aRNA-LongSAGE protocol,

the *in vitro* transcription was carried out for 18 h. Shorter incubation times are possible, especially if more than 40 ng of total RNA are available for the amplification procedure. The yield of aRNA can be estimated using an RNA PicoChip on a BioAnalyzer platform (Agilent, Böblingen, Germany). A minimum of 1.2 µg of aRNA should be used for the generation of an aRNA-longSAGE library (*see* **Note 3**).

3.2. cDNA Synthesis

1. Add DEPC-treated water to the aRNA to a final volume of 10 µL, then add 2 µL of SAGErandom oligonucleotide and 1 µL 10 m*M* dNTP mix and incubate for 5 min at 65 °C in a thermocycler. After the incubation place sample on dry ice, thaw on wet ice and add 4 µL First Strand Buffer, 1 µL 0.1 *M* DTT, 1 µL RNaseOUT, and 1 µL SuperScript™ III Reverse Transcriptase. Incubate in a thermocycler for 5 min at 37 °C, 1 h at 50 °C, and 15 min at 70 °C.
2. Add 1 µL of RNase H (2 U/µL, diluted in 1X Second Strand Buffer) and incubate 20 min at 37 °C in a thermocycler.
3. Add 0.5 µL DEPC-treated water, mix well, and remove 0.5 µL of the sample for loading on a 1 % agarose gel.
4. Add 79 µL of DEPC-treated water.
5. Wash 200 µL Oligo$(dT)_{25}$ Beads with 100 µL of Binding Buffer and resuspend in 100 µL of Binding Buffer.
6. Mix sample with resuspended beads in a siliconized microcentrifuge tube **(Note 4)**. Put sample in an overhead shaker and rotate for 15 min at room temperature.
7. Wash sample twice with 200 µL Washing Buffer B and four times with 200 µL Second Strand Buffer.
8. Resuspend beads in 112.25 µL of icecold DEPC-treated water and add the following components on ice: 32 µL 5X Second Strand Buffer, 6 µL 0.1 *M* DTT, 3 µL dNTP-Mix (10 m*M* each), 4.5 µL *E. coli* DNA Polymerase I (11.8 U/µL), 1.5 µL *E. coli* DNA Ligase (10 U/µL) and 0,75 µL *E. coli* RNaseH (2 U/µL, dilution in Second Strand Buffer).
9. Incubate in a thermomixer for 2.5 h at 16 °C. To keep beads in suspension, mix sample every 15 min on a vortexer at slow speed (use a setting of 5).
10. Add 4 µL T4 DNA Polymerase (3 U/µL) and incubate for 5 min at 16 °C.
11. Add 4 µL 0.5 *M* EDTA and 750 µL 1X BW and incubate for 20 min at 75 °C.
12. Wash beads once with 750 µL 1X BW, four times with 750 µL 1X BW/1X BSA and twice with 200 µL 1X Promega Buffer K that contains 0.1 mg/mL BSA.

3.3. Cleavage of cDNA With the Anchoring Enzyme (Hsp92 II)

1. Resuspend beads in 200 µL reaction mix containing 1X Buffer K (Promega), 0.1 mg/mL BSA, and 50 U Hsp92II (*see* **Note 2**) and incubate in a thermomixer for 1 h at 37 °C. To keep beads in suspension, mix sample every 15 min on a vortexer at slow speed (use a setting of 5).
2. Wash beads once with 750 µL 1X BW, four times with 750 µL 1X BW/1X BSA, and twice with 200 µL 1X Ligase Buffer.
3. Resuspend beads in 200 µL 1X Ligase Buffer.

3.4. Ligating Linkers to Bound cDNA

Prior to the first use the linker oligonucleotides are phosphorylated and hybridized to obtain linkers 1 and 2. Phosphorylated and hybridized linkers can be stored at −20 °C in aliquots for single use.

1. Linker oligonucleotides 1B and 2B are phosphorylated in two separate tubes by adding 6 µL LoTE, 2 µL 10X polynucleotide kinase buffer, 2 µL 10 m*M* ATP and 1 µL T4 polynucleotide kinase (10 U/µL) to 9 µL of Linker oligonucleotide (350 ng/µL). The tubes are incubated in a thermocycler for 30 min at 37 °C and then for 10 min at 65 °C.
2. To hybridize linkers mix the phosphorylated Linker B molecules with 9 µL of the appropriate Linker A oligonucleotide (350 ng/µL), i.e., mix phosphorylated Linker 1B with 9 µL Linker 1A and phosphorylated Linker 2B with 9 µL Linker 2A. Incubate both tubes for 2 min at 95 °C, 10 minutes at 65 °C, 10 min at 37 °C and 20 min at 22 °C in a thermocycler. Add 271 µL of LoTE to each tube, aliquot and store linkers at −20 °C (final concentration of linkers is 20 ng/µL).
3. To ligate linkers to the immobilized cDNA divide sample (200 µL beads in 1X Ligase Buffer) equally in two new tubes.
4. Remove the supernatant and resuspend in 9 µL reaction mix containing 5 µL LoTE, 2 µL 5X Ligase Buffer and 2 µL of phosphorylated and annealed Linker 1 or 2, respectively.
5. Incubate sample for 2 min at 50 °C then for 10 min at room temperature.
6. Add 1 µL HC T4 Ligase (5 U/µL) to each tube, and vortex carefully
7. Incubate at 16 °C for 1.75 h in a thermomixer. To keep beads in suspension, mix sample every 15 min on a vortexer at slow speed (use a setting of 5).
8. Wash beads once with 500 µL 1X BW/1X BSA.
9. Pool ligation reactions 1 and 2 in a new tube.
10. Wash beads three times with 500 µL 1X BW/1X BSA, once with 200 µL 1X BW, and once with 200 µL 1X Buffer 4 (NEB).
11. Resuspend beads in 200 µL 1X Buffer 4 (NEB) and store overnight at 4 °C.
12. Wash beads twice with 200 µL 1X Buffer 4 prewarmed to 37 °C.

3.5. Release of cDNA Tags Using the Tagging Enzyme MmeI

1. Prepare a 1 m*M* SAM solution by diluting the 32 m*M* SAM solution that comes with the *Mme*I enzyme.
2. Resuspend beads in 200 μL prewarmed (37 °C) reaction mix containing 1X Buffer 4 (NEB), 0.05 m*M* SAM, and 8 U of *Mme*I.
3. Incubate at 37 °C for 1 h in a thermomixer. To keep beads in suspension, gently mix sample every 15 min on a vortexer (use setting 5).
4. Centrifuge at 16,110×g for 2 min in a microcentrifuge.
5. Transfer supernatant to a new microcentrifuge tube (no need to use siliconized tubes any longer).
6. Resuspend beads in 40 μL LoTE.
7. Centrifuge at 16,110×g for 2 min.
8. Remove supernatant and pool it with the supernatant of the first centrifugation step (total volume: 240 μL)
9. Extract with PC8: add an equal volume of PC8 to the sample, mix well on a vortex, centrifuge at 16,110×g for 2 min, and transfer the upper (aqueous) phase to a fresh microcentrifuge tube.
10. Transfer 40 μL of the sample to a new as the "no ligase" control during ditag ligation and PCR amplification of ditags. Dilute this negative control with 160 μL of LoTE
11. Precipitate both sample and negative control by adding 100 μL 10 *M* ammonium acetate, 3 μL glycogen and 1 mL 100 % ethanol and centrifuge 30 min at 4 °C and 16,110×g.
12. Wash each pellet three times with 500 μL of 70 % ethanol.
13. Resuspend sample in 1.5 μL of LoTE and 2.5 μL of water; resuspend negative control in 1.5 μL of LoTE and 3.3 μL of water. Incubate both tubes at room temperature for 5 min.

3.6. Ligating Tags to Form Ditags

1. Add 1.2 μL 5X Ligase Buffer to sample and negative control, then add 0.8 μL HC T4 Ligase (5 U/μL) to sample but not to negative control.
2. Incubate for 2.5 h at 16 °C in a thermocycler.
3. Add 15 μL LoTE to sample and to negative control.

3.7. PCR Amplification of Ditags

To optimize PCR conditions a test PCR is run using different dilutions of the ditags (1:50/1:100/1:200/1:400 in LoTE) as a template at 26, 28, and 31 PCR cycles. A 1:50 dilution of the minus-ligase control run at 31 cycles serves as a negative control. Prepare the PCR reactions in a laminar-flow hood to avoid contamination of the sample.

1. For each PCR reaction, mix 1 μL of template (diluted ditags), 4 μL of 10X BV-Mg Buffer, 3 μL of DMSO, 5 μL of dNTP mix (10 m*M* each), 1 μL of each PCR primer (350 ng/μL) and 25 μL of water.
2. Add a drop of mineral oil to each well and incubate in a thermocycler for 3 min at 95 °C then hold temperature at 78 °C and add 10 μL of polymerase mix containing 3 μL of Taq polymerase in 1X BV-Mg Buffer to each well.
3. Run PCR for 26, 28, and 31 cycles in parallel, each cycle consisting of 30 s at 95 °C, 30 s at 55 °C, and 30 s at 70 °C. After the last PCR cycle, incubate for 5 min at 70 °C.
4. Load 5 μL of each PCR reaction on a 20 × 20 cm polyacrylamide gel (12 %, 19:1 acrylamide/bis). Run gel in TAE buffer at 180 V until the bromphenol blue band of the marker has travelled a distance of about 8 cm (*see* **Note 5**). Stain gel in SYBR Green I solution for 15 min and make bands visible under ultraviolet (UV) light. Set up a large-scale PCR of 96 PCR reactions using the conditions that were determined as optimal earlier in this section. Following amplification, pool all reactions in a 15-mL polypropylene tube (*see* **Note 6**).
5. Centrifuge for 1 min at 2630×g and remove the mineral oil.
6. Extract with an equal volume of PC8, then centrifuge for 10 min at 2200×g.
7. Transfer 2.1 mL of the sample (upper phase) in each of two centrifuge tubes, add 700 μL of 10 *M* ammonium acetate, 18 μL of glycogen and 6 mL of ethanol to each tube. Mix well and centrifuge for 30 min at 4 °C and 12,000×g.
8. Wash pellet twice with 5 mL 70 % ethanol, remove supernatant, and air-dry the pellet.
9. Resuspend each pellet in 45 μL of LoTE and incubate for 5 to 10 min at 37 °C to aid solubilization.

3.8. Isolation of Ditags

1. Mix the complete sample (approx 90 μL) with 68 μL of water, 20 μL of 10X Buffer K, 2 μL of BSA (10 mg/mL), and 20 μL of Hsp92II (10 U/μL) and incubate for 1 h at 37 °C in a heating block.
2. Extract with PC8, then add 66.7 μL of 10 *M* ammonium acetate, 3 μL of glycogen, and 1 mL of ethanol and precipitate the ditags overnight at −70 °C.
3. Centrifuge for 30 min at 4 °C and 16,110×g wash the pellet twice with 500 μL of 70 % ethanol, dry pellet for 10 min at 16 °C, and resuspend the pellet in 90 μL of TE.
4. Add 5 μL of glycerol (*see* **Note 7**).
5. Load complete sample on a 20 × 20 cm polyacrylamide gel (12 %, 19:1 acrylamide/bis). Run gel at 4 °C and 180 V until the bromphenol blue band of the marker has travelled a distance of about 8 cm (*see* **Note 5**). Stain gel in SYBR Green I solution for 15 min and make bands visible under UV light.
6. Cut out the 34-bp ditag band.

7. For electroelution of the ditags, prepare the electroelution device in such a way that the elution chamber is 2 U-inserts wide and the trap is 1 U-insert wide. Put the gel slices into the elution chamber and electroelute at 4 °C and 150 V for 2 h, then reverse polarity and turn on 200 V for 20 s. Transfer the eluted ditags (1 mL sample volume) from the trap to two microcentrifuge tubes (*see* **Note 8**).
8. Extract with PC8.
9. Extract the aqueous phase with an equal volume of chloroform.
10. Precipitate ditags by adding 50 µL of 3 *M* sodium acetate, 2 µL glycogen, and 1250 µL ethanol to each tube. Incubate at −70 °C overnight.
11. Centrifuge at 4 °C and 16,110×g for 30 min, wash the pellets twice with 500 µL of 70 % ethanol, air-dry pellets on ice and resuspend both pellets in a total volume of 7 µL of LoTE.

3.9. Concatenation of Ditags

1. Add 2 µL of 5X Ligase Buffer and 1 µL of T4 Ligase HC (5 U/µL) and incubate in a thermocycler at 16 °C for 30 min.
2. Add 190 µL of LoTE and extract with 200 µL PC8. Transfer the supernatant to a new tube.
3. Add 100 µL of 10 *M* ammonium acetate solution, 3 µL of glycogen, and 700 µL of ethanol, keep on ice for 10 min, then centrifuge for 15 min at 16,110×g. Wash pellet twice with 500 µL of 70 % ethanol and resuspend in 10 µL of LoTE.
4. Add 5 µL of loading buffer, incubate for 10 min at 65 °C, chill sample on ice, and load on a single lane of an 8 % polyacrylamide gel (acrylamide/bis 19:1; *see* **Note 9**).
5. Electrophorese for 3 h at 130 V. Stain gel in SYBR Green I solution for 15 min.
6. Visualize the bands under UV light and excise concatemers >300 bp from the gel (*see* **Note 10**). Do not excise the large concatemers at the upper edge of the well (leave a margin of 1 mm gel at the upper edge of the well).
7. For electroelution of the concatemers, prepare the electroelution device in such a way that the elution chamber is 2 U-inserts wide and the trap is 1 U-insert wide. Put the gel slices into the elution chamber and electroelute for 60 min at room temperature, then reverse polarity and turn on 200 V for 20 s. Transfer the eluted ditags (1 mL sample volume) from the trap to two microcentrifuge tubes.
8. Extract with PC8.
9. Extract the aqueous phase with an equal volume of chloroform.
10. Precipitate ditags by adding 50 µL of 3 *M* sodium acetate, 2 µL glycogen and 1250 µL ethanol to each tube. Incubate at −20 °C for 1 h or overnight.
11. Centrifuge at 4 °C and 16,110×g for 15 min, wash the pellets twice with 500 µL of 70 % ethanol, air-dry pellets and resuspend both pellets in a total volume of 15 µL of water.

3.10. Cloning of Concatemers

1. Mix 1 µL of pZErO-1 Supercoiled cloning vector (1 µg/µL) with 2 µL of 10x Buffer 2 (NEB), 16 µL of water, and 1 µL of *Sph*I (5 U/µL).
2. Incubate for 15 min at 37 °C in a water bath (*see* **Note 11**).
3. Add 180 µL of loTE and extract with 200 µL of PC8.
4. Precipitate the linearized vector by adding 66.7 µL of 10 *M* ammonium acetate, 3 µL of glycogen and 1 mL of ethanol and centrifuging for 10 min at 16,110×g.
5. Wash pellet three times with 500 µL of 70 % ethanol.
6. Resuspend the air-dried pellet in 40 µL of TE (final concentration of the linearized vector is 25 ng/µL).
7. Mix 1 µL of the linearized pZErO-1 with 6 µL of concatemers, 2 µL of 5X Ligase Buffer, and 1 µL T4 DNA ligase (1 U/µL). Incubate for 1 h at 16 °C and for another hour at room temperature.
8. Add 190 µL of LoTE and do a PC8 extraction with 200 µL of PC8.
9. Precipitate the sample by adding 66.7 µL of 10 *M* ammonium acetate, 3 µL of glycogen and 1 mL of ethanol and centrifuging for 20 min at 16,110×g and 4 °C.
10. Wash pellet four times with 500 µL of 70 % ethanol.
11. Resuspend the air-dried pellet in 8 µL of LoTE

3.11. Transformation of Bacteria

1. Use 0.8 µL of the ligation product to electroporate an aliquot (40 µL) of electrocompetent *E. coli* Top 10 (Voltage: 1800 V; *see* **Note 12**). Resuspend electroporated bacteria in 1 mL of LB medium and incubate for 1 h at 37 °C shaking at 220 rpm.
2. Plate 300 µL bacteria suspension on each of three 14.5 cm LB-Zeocin-X-Gal plates.

3.12. Insert-PCR

1. For each PCR reaction, mix 2 µL of 5X RDA-Buffer, 1.2 µL of 50 m*M* $MgCl_2$, 0.3 µL of each primer, 0.3 µL of 10 m*M* dNTP-Mix, and 5.9 µL of water.
2. Pipet 10 µL of the PCR-mix into the wells of a 96-well plate and add a drop of mineral oil into each well.
3. Use a sterile toothpick to gently touch a white bacteria colony (*see* **Note 13**), and then dip it into the PCR mix.
4. To each well and incubate in a thermocycler for 2 min at 95 °C, then hold temperature at 78 °C and add 5 µL of polymerase mix containing 1 µL of Taq polymerase in 1X RDA Buffer to each well. Run five cycles consisting of 30 s at 95 °C, 30 s at 60 °C and 45 s at 72 °C, and then run additional 30 cycles consisting of 30 s at 95 °C and 60 s at 70 °C.
5. Run 5 µL of each PCR reaction of a 1.5 % agarose gel to check the insert sizes of the SAGE library. Empty vectors will give a 330-bp PCR product.

4. Notes

1. Unless stated otherwise, water means water with a conductivity of at least 18 MΩ.
2. Hsp92II is an isoschizomer of *Nla*III that can be stored at −20 °C. As a result of different unit definitions for Hsp92II and *Nla*III, the volume of Hsp92II that is needed for digestion steps is much higher than the volume of *Nla*III.
3. If there is more than 1.2 μg of aRNA available, use up to 2.5 μg of aRNA for the generation of an aRNA-LongSAGE library. More starting material tends to generate larger insert sizes in our hands.
4. Use siliconized tubes when dealing with magnetic beads to prevent beads from adsorbing to the surface of the tube. Wash beads by resuspending on a vortexer at slow speed (use a setting of 5) instead of pipetting the beads up and down, in order to minimize loss of beads by adsorption to pipet tips.
5. Keeping the travelling distance constant will result in equal electrophoresis conditions between libraries and is better than keeping travelling time constant. Eight centimeters of travelling distance on a 20 × 20 cm gel gives a good separation of ditags from linkers.
6. Make sure not to use polystyrene tubes, as polystyrene reacts with PC8.
7. Glycerol is added instead of loading buffer to avoid contamination of the ditags. Adding glycerol is essential in order to increase the density of the sample. Without glycerol, the sample will diffuse into the running buffer and be lost.
8. Electroelution from the polyacrylamide gel gives a higher yield of DNA than elution by diffusion as stipulated in the standard SAGE protocol.
9. This is a different gel from that used in the standard SAGE protocol. Using an acrylamide/bis proportion of 19:1 instead of 37.5:1 better separates undesired small concatemers from the concatemers that are cut out from the gel.
10. For no obvious reason, the concatenation step may not work on the first try for each library. As this protocol does use only a small fraction of synthesized ditags as template for large-scale PCR, it is possible to try again with a new large-scale PCR.
11. A fully linearized vector is important for the success of the cloning step. Check on an agarose gel whether the vector is fully linearized. If the digestion with *Sph*I was not complete, the digestion should be repeated with a longer incubation time or with more than 5 U of *Sph*I.
12. Use bacteria from the Zero Background Cloning Kit (Invitrogen). Prepare competent bacteria according to the instructions given in the kit. In our hands, this bacteria strain is better than the *E. coli* DH10B recommended in the original SAGE protocol.
13. Blue–white screening helps to chose colonies with large inserts. Even though there are white colonies with short inserts as well as blue colonies with long inserts, all in all, the average insert size is longer for white colonies than for blue ones.

References

1. Heidenblut, A. M., Luttges, J., Buchholz, M., et al. (2004) aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Res*. **32,** E131.
2. Van Gelder, R. N., von Zastrow, M. E., Yool, A., Dement, W. C., Barchas, J. D. and Eberwine, J. H. (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87,** 1663–1667.
3. Puskas, L. G., Zvara, A., Hackler, L., Jr. and Van Hummelen, P. (2002) RNA amplification results in reproducible microarray data with slight ratio bias. *Biotechniques* **32,** 1330–1334, 1336, 1338, 1340.
4. Polacek, D. C., Passerini, A. G., Shi, C., et al. (2003) Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. *Physiol. Genomics* **13,** 147–156.
5. Feldman, A. L., Costouros, N. G., Wang, E., et al. (2002) Advantages of mRNA amplification for microarray analysis. *Biotechniques* **33,** 906–912, 914.
6. St Croix, B., Rago, C., Velculescu, V., et al. (2000) Genes expressed in human tumor endothelium. *Science* **289,** 1197–1202.
7. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **20,** 508–512.

4

SuperSAGE

Hideo Matsumura, Monika Reuter, Detlev H. Krüger, Peter Winter, Günter Kahl, and Ryohei Terauchi

Summary

As a tool for high-throughput, quantitative gene expression analysis, serial analysis of gene expression (SAGE) is one of the most powerful techniques. However, the short size of tags (14 bp) has hindered the application of SAGE to a vast majority of eukaryotes without sufficient genomic resources, including expressed sequence tag and genome sequences. To overcome this problem, we developed SuperSAGE, which is based on 26-bp tags from complementary DNA (cDNA), using EcoP15I as a tagging enzyme. Because longer cDNA fragments can easily be recovered by 3′-rapid amplification of cDNA ends (RACE) PCR using primers corresponding to the 26-bp tag sequences in non-model organisms, SuperSAGE allows the identification of novel genes in all eukaryotic organisms, and recommends itself as a useful platform in various fields of biological studies.

Here, we present an updated SuperSAGE protocol, which incorporates several modifications and some recommendations to avoid total failure, particularly in the EcoP15I digestion step.

Key Words: SuperSAGE; EcoP15I; non-model organisms; transcript profiling; expression markers.

1. Introduction

The concept of so-called expression markers evolved only a few years ago, although expressed sequences were highly praised for a full decade. For example, complementary DNAs (cDNAs) or expressed sequence tags (ESTs), short synthetic oligonucleotides of 300–500 bp, complementary to the 5′ or 3′ end of a specific messenger RNA and usually derived from a cDNA library by

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

random sequencing, represent tags for the state of gene expression in a cell or tissue type at a given time. Both types have been generated by the hundreds of thousands and used to design expression arrays, high-density chips onto which multiple cDNAs, short fragments of cDNAs, ESTs, or gene fragments are spotted, that allow to determine the expression of multiple genes simultaneously. For an expression assay, labeled cDNA from a target tissue is hybridized to the expression array, and hybridization patterns are directly converted into information about expressed genes in the sample. Such arrays are now standard tools to generate expression profiles (transcript profiles, expression fingerprints), complex, context-dependent and genome-wide patterns of (preferably all) expressed genes at a given time. The expression profile is characteristic for a certain cell, tissue, organ, or organisms (e.g., a bacterial cell), but changes continuously, depending on developmental stage and environment. Although in worldwide use, expression microarrays still suffer from platform- or experiment-specific problems ***(1,2)***. The major drawback for all microarray-based expression analyses, however, resides in the fact that expression can be monitored only of those genes whose sequences are on the chip ("closed architecture" format). Therefore, gene discovery is not possible with this approach.

More recently, other techniques that produce expression markers in a high-throughput, genome-wide format without this drawback appeared. Among these, real competitive (rc)PCR ***(3)***, massively parallel signature sequencing (MPSS) ***(4)***, and serial analysis of gene expression (SAGE) ***(5)*** stand out because of their potential and widespread use. In spite of rather complex protocols, these techniques are quantitative (i.e., the number of each transcript at any time point can be estimated), comprehensive (i.e., all transcripts can be recovered), and informative (i.e., each tag extracted from a messenger RNA can be annotated with high fidelity). It is for these (and other) reasons that we favor these "open architecture" platforms. Here, we restrict ourselves to a detailed protocol of a substantially improved variant of the conventional SAGE procedure, coined SuperSAGE ***(6–8)***. The generation of expression markers with this technology delivers identifiers (tags) of 26 bp for each and every cDNA, the longest tags ever produced in the SAGE family, which allow unequivocal and secure annotation to the underlying genes. More excitingly, SuperSAGE, as the only technique available, permits the simultaneous detection and quantitation of the transcriptome from two intimately interacting organisms (e.g., a host and a pathogen, commensal, or parasite) without the physical separation of both. SuperSAGE tags are also superior agents for RNA interference (RNAi), because RNA duplexes of 27–29 bp length can be 100-fold more potent than the traditional 21-mer small interfering RNAs (siRNAs). The

enhanced potency of the longer duplexes reflects their higher affinity to Dicer, which links the production of siRNAs to their incorporation into the RNA-induced silencing complex (RISC) ***(9,10)***. Last but not least, SuperSAGE tags have been spotted directly onto microarrays and were successfully hybridized to various cDNA samples (Matsumura et al., 2006).

We expect that the increasing use of SuperSAGE, with all of its advantages, will add to our technical repertoire for transcript profiling and to our knowledge of the transcriptome and its changes in the life cycle of an organism.

2. Materials

2.1. Linker Preparation

1. Linker oligonucleotides: end-labeled linker oligonucleotides are synthesized by Operon Biotechnologies, Japan:
   Linker-1EA:
   (FITC5′–TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATACAGCA
   GCATG–3′)
   Linker-1EB:
   (5′-CTGCTGTATTAAGCCTAGTTGTACTGCACCAGCAAATCCAAA-
   3′Amino), Linker-2EA:
   (FITC5′–TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGCAGCAG
   CATG–3′), Linker-2EB:
   (5′-CTGCTGCGTACATCGTTAGAAGCTTGAATTCGAGCAGAAA-3′amino).
   These oligonucleotides are purified by high-performance liquid chromatography (HPLC) (Operon Biotechnologies, Japan).
2. LoTE buffer: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* ethylenediamine tetraacetic acid (EDTA).
3. Polynucleotide kinase buffer (10X): 0.5 *M* Tris-HCl, pH 8.0, 0.1 *M* $MgCl_2$ and 50 m*M* dithiothreitol (DTT) (Takara, Japan).
4. T4 polynucleotide kinase (10 U/μL): Store at −20 °C (Takara, Japan).
5. ATP solution: prepare 10 m*M* solution by dilution of a 100 m*M* ATP-lithium stock (Roche Diagnostics, Germany).

2.2. Double-Strand cDNA Synthesis

1. First strand buffer (5X): 250 m*M* Tris-HCl, pH 8.0, 375 m*M* KCl and 15 m*M* $MgCl_2$(Invitrogen, Carlsbad, CA).
2. Biotinylated adapter-oligo (dT) primer: dissolve synthesized biotin-labeled oligonucleotides (5′-biotin-CTGATCTAGAGGTACCGGATCCCAGCAGTTTT
   TTTTTTTTTTTTT-3′) in LoTE (1 μg/μL). This oligonucleotide was HPLC-purified by the producer (Operon Biotechnologies, Japan).
3. 0.1 *M* DTT (Invitrogen, Carlsbad, CA).

4. 10 m*M* dNTP: 10 m*M* each of dATP, dTTP, dCTP, and dGTP (Invitrogen, Carlsbad, CA).
5. Cloned M-MLV reverse transcriptase (Invitrogen, Carlsbad, CA).
6. Second strand buffer: 188 m*M* Tris-HCl, pH8.3, 906 m*M* KCl, 100 m*M* $(NH_4)_2SO_4$, 46 m*M* $MgCl_2$, 37.5 m*M* DTT and 1.5 m*M* NAD (Invitrogen, Carlsbad, CA).
7. *Escherichia coli* DNA polymerase (10 U/μL, Invitrogen, Carlsbad, CA).
8. *E. coli* DNA ligase (1.2 U/μL, Invitrogen, Carlsbad, CA).
9. *E. coli* RNase H (2 U/μL, Invitrogen, Carlsbad, CA).
10. Phenol:chloroform:isoamyl alcohol (25:24:1) (Invitrogen, Carlsbad, CA)
11. Ammonium acetate: 10 *M* solution (Wako Chemicals, Japan). Store at room temperature.
12. Glycogen solution (20 mg/mL; Roche Diagnostics, Germany).

2.3. 26-bp Tag Extraction From cDNA.

1. NlaIII (10 U/μL): Store at −70 °C (New England Biolabs, Ipswich, MA).
2. NlaIII digestion buffer (NEBuffer 4) (10X): 20 m*M* Tris-acetate, pH 7.9, 50 m*M* potassium acetate, 10 m*M* magnesium acetate and 1 m*M* DTT (New England Biolabs, Ipswich, MA).
3. Bovine serum albumin (BSA) (10 mg/mL) (New England Biolabs, Ipswich, MA).
4. Streptavidin-coated magnetic beads (Streptavidin MagneSphere Paramagnetic Particles) (1 mg/mL): store at 4 °C (*see* **Note 1**) (Promega, Madison, WI).
5. Binding and washing (B&W) buffer (2X): 10 m*M* Tris-HCl, pH 7.5, 1 m*M* EDTA and 2 *M* NaCl. Store at room temperature.
6. T4-DNA ligase (2000 U/μL): store at −20 °C (New England Biolabs, Ipswich, MA).
7. T4-DNA ligase buffer (5X): 250 m*M* Tris-HCl, pH 7.5, 50 m*M* $MgCl_2$, 5 m*M* ATP, 50 m*M* DTT and 125 μg/mL BSA (New England Biolabs, Ipswich, MA).
8. EcoP15I (2 U/μL): store at −20 °C (*see* **Note 2**) (New England Biolabs, Ipswich, MA).
9. EcoP15I digestion buffer (10X): 100 m*M* Tris-HCl, pH 8.0, 100 m*M* KCl, 100 m*M* $MgCl_2$, 1 m*M* EDTA, 1 m*M* DTT and 50 μg/mL BSA. Store at −20 °C (*see* **Note 3**).

2.4. Purification of 26-bp Tags

1. Acrylamide/BIS solution (40 %, 19:1): Store at 4 °C (SERVA, Germany).
2. *N,N,N,N*'-tetramethyl-ethylenediamine (TEMED) (Wako Chemicals, Japan).
3. Ammonium persulfate: prepare 10 % solution in sterilized water and store at 4 °C.
4. 6X loading dye: 30 %(v/v) glycerol, 0.25 % (w/v) bromophenol blue, and 0.25 % (w/v) xylene cyanol.
5. SYBR green solution: dilute original SYBR green stock solution 10,000 times with 1X TAE buffer (Molecular Probe, Eugene, OR). Store at 4 °C.
6. Spin-X® column (Corning, Corning, NY).

2.5. Ditag Formation

1. KOD DNA polymerase in Blunting high kit (TOYOBO, Japan).
2. Blunting buffer (10X) in Blunting high kit (TOYOBO, Japan).
3. Ligation high solution in Blunting high kit (TOYOBO, Japan).

2.6. Ditag PCR

1. GeneAmp 10X PCR Gold Buffer (Applied Biosystems, Foster City, CA).
2. dNTP solution: 2 m*M* each of dATP, dTTP, dCTP, and dGTP (Applied Biosystems, Foster City, CA).
3. 25 m*M* $MgCl_2$ solution. (Applied Biosystems, Foster City, CA).
4. Biotinylated ditag amplification primers: synthesize biotinylated oligonucleotides (Ditag1E: 5′-biotin-CAACTAGGCTTAATACAGCAGCA-3′, Ditag2E: 5′-biotin-CTAACGATGTACGCAGCAGCA-3′) by Operon Biotechnologies, Japan and prepare a solution of 350 ng/μL from each.
5. AmpliTaq Gold (5 U/μL): store at −20 °C (Applied Biosystems, Foster City, CA).

2.7. Purification of Ditag PCR Products and Second NlaIII Digestion

1. Binding buffer (PB buffer) in Qiaquick PCR purification kit (Qiagen, Germany).
2. Qiaquick spin column in Qiaquick PCR purification kit (Qiagen, Germany).
3. Washing buffer (PE buffer, 5X) (Qiagen, Germany): prepare 1X solution by adding ethanol before use.

2.8. Concatemer Formation and Purification

1. Buffer ERC in MinElute Reaction Cleanup Kit (Qiagen, Germany).
2. MinElute Spin Column in MinElute Reaction Cleanup Kit (Qiagen, Germany).

2.9. Vector Cloning and Transformation

1. pGEM-3Z, plasmid cloning vector (Promega, Madison, WI).
2. SphI (5 U/μL): store at −20 °C. (New England Biolabs, Ipswich, MA).
3. SphI digestion buffer (NEBuffer 2) (10X): 10 m*M* Tris-HCl, pH 7.9, 10 m*M* $MgCl_2$, 50 m*M* NaCl, and 1 m*M* DTT (New England Biolabs, Ipswich, MA).
4. Calf intestine alkaline phosphatase (CIAP) (10 U/μL) (Takara, Japan).
5. Alkaline phosphatase buffer (10X): 500 m*M* Tris-HCl, pH 9.0 and 10 m*M* $MgCl_2$ (Takara, Japan).
6. Electro-competent *E. coli* cells (ElectroMAX DH10B): store at −80 °C (Invitrogen, Carlsbad, CA).
7. Liquid culture medium (SOC medium): 2 % tryptone, 0.5 % yeast extract, 10 *M* NaCl, 2.5 m*M* KCl, 10 m*M* $MgCl_2$, 10 m*M* $MgSO_4$ and 20 m*M* glucose (Invitrogen, Carlsbad, CA).

8. *E. coli* plate medium I (Luria-Bertani [LB] medium): 1 % tryptone, 0.5 % yeast extracts, 1 % NaCl, 1.5 % agar, containing ampicillin (100 μg/mL), 0.004 % 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal), and 0.1 m*M* isopropyl-β-D-thiogalactopyranoside (IPTG).

2.10. Colony PCR and Sequencing

1. Taq polymerase (TAKARA Taq, 5 U/μL) (Takara, Japan).
2. PCR buffer (10X): 100 m*M* Tris-HCl, pH 8.3, 500 m*M* KCl and 15 m*M* $MgCl_2$ (Takara, Japan).
3. dNTP solution: 2.5 m*M* each of dATP, dTTP, dCTP, and dGTP (Takara, Japan).
4. PCR primers for amplification of inserted fragments in pGEM3Z: M13F: 5′- GTAAAACGACGGCCAGT-3′, M13RV: 5′-GGAAACAGCTATGACCATG-3′ (Operon Biotechnologies, Japan).

3. Methods

In SuperSAGE, tag extraction from cDNA and the corresponding steps are modified, so that 26-bp tags are obtained by EcoP15I digestion, whereas the basic experimental procedure principally follows the original SAGE protocol.

Although a substantially improved EcoP15I production procedure that yields a highly concentrated and pure EcoP15I batch has been developed ***(12)***, we present a SuperSAGE protocol using a commercially supplied enzyme (*see* **Note 2**). However, as stated in **Note 3**, it is imperative that EcoP15I digestion buffer be used as described in the present protocol to obtain optimum results.

Compared to the original SAGE method, an additional purification step (polyacrylamide gel electrophoresis [PAGE] purification) is required after EcoP15I digestion, because EcoP15I also cuts cDNA sites adjacent to the poly-A tail, and fragments of various sizes (usually longer than linker-26 bp tag fragments) are also released from the beads. This purification additionally eliminates most of the linkers and linker dimers, which compete with ditags in PCR amplification, from the solution. With this modification, we can reliably and efficiently amplify ditags from diluted templates **(Fig. 1)**. This modification should be applicable to other SAGE or LongSAGE methods to increase the recovery of ditags after PCR.

3.1. Linker Preparation

1. Dissolve synthesized linker oligonucleotides (Linker-1EA, -1EB, -2EA, -2EB) in LoTE buffer (1 μg/μL).
2. Mix1 μL Linker-1EB or Linker-2EB, 1 μL 10X polynucleotide kinase buffer, 1 μL 10 m*M* ATP, 10 μL H_2O, and 1 μL T4 polynucleotide kinase for phosphorylation of linker oligonucleotides, and incubate at 37 °C for 30 min.

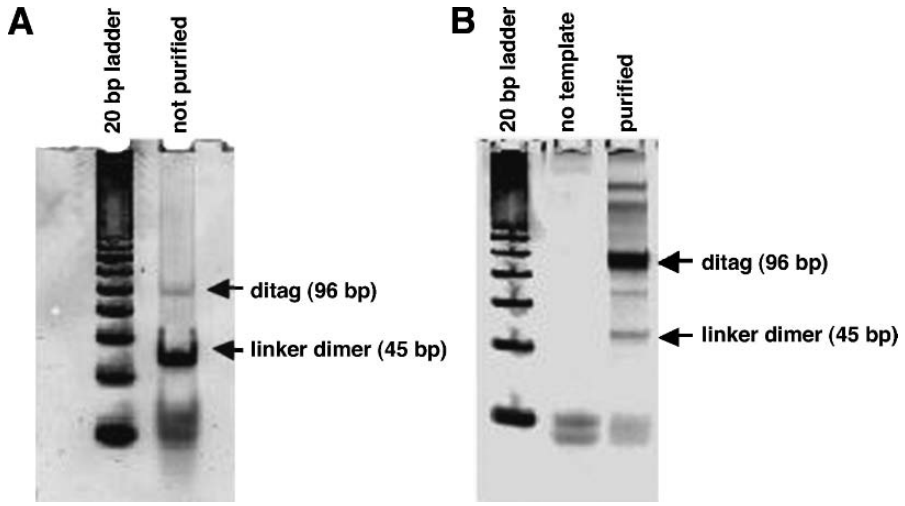


Fig. 1. Purification of linker-26 bp tag fragments enhances the efficiency of PCR amplification of 96-bp ditags. **(A)** Ligation product, which was not purified with polyacrylamide gel electrophoresis after EcoP15I digestion, was used as template for ditag PCR amplification. **(B)** Ligation product of purified linker-26 bp tag fragment was used as template for ditag PCR amplification. Arrows indicate amplified ditag fragment (96 bp) and linker dimer fragment (45 bp). The lane "no template" in **B** represents the PCR reaction product without template DNA (blank control).

3. Add 1 μL Linker–1EA or –2EA to the 5′-phosphorylated Linker-1EB or –2EB solution from the previous step, respectively. After mixing, denature by incubating at 95 °C for 2 min and cool down to 20 °C for annealing Linker–1EA and –1EB, and Linker-2EA, and –2EB. The annealed double-stranded DNAs are designated Linker-1E and Linker-2E, respectively.

3.2. Double-Stranded cDNA Synthesis

1. The synthesis of double-stranded cDNA was performed as described in the cDNA Synthesis System (Invitrogen), except for the oligo (dT) primer. Dissolve purified poly(A)$^+$ RNA (2–5 μg) in 29 μL diethylpyrocarbonate (DEPC)-treated water. For first strand cDNA synthesis, add 10 μL 5X first strand buffer, 1 μL biotinylated adapter-oligo dT primer, 5 μL 0.1 *M* DTT, 2.5 μL 10 m*M* dNTP, and 2.5 μL M-MLV reverse transcriptase to poly(A)$^+$ RNA, and incubate at 37 °C for 1 h.
2. For second strand cDNA synthesis, add 40 μL 10X second strand buffer, 295 μL DEPC-treated water, 7.5 μL 10 m*M* dNTP, 10 μL *E. coli* DNA polymerase, 1.75 μL *E. coli* RNase H, and 1.25 μL *E. coli* DNA ligase to 45 μL first strand cDNA solution, and mix. Incubate at 16 °C for 2 h.

3. Run 10 μL second strand cDNA aliquot, taken out of the reaction solution, on a 1 % TAE-agarose gel. The sizes of most of double-stranded cDNAs should range from 0.5 to 2.0 kbp (*see* **Note 4**).
4. Add half a volume of phenol:chloroform:isoamyl alcohol (195 μL) to the remaining second strand cDNA solution (390 μL), vortex briefly, and spin at 10,000 g for a few minutes. Transfer the upper aqueous layer to a new tube (*see* **Note 5**). For ethanol precipitation of cDNA, add 100 μL 10 *M* ammonium acetate, 3 μL glycogen, and 1000 μL cold ethanol to the phenol/chloroform extracted solution and keep at −80 °C for 1 h. Centrifuge at 15,000 g at 4 °C for 30 min. Wash the pellet twice with 70 % ethanol, and dry. Dissolve the precipitated cDNA in 20 μL LoTE buffer.

3.3. 26-bp Tag Extraction From cDNA

1. Add 20 μL NlaIII digestion buffer (NEBuffer 4), 2 μL BSA, 152 μL LoTE, and 5 μL NlaIII to the cDNA solution, mix, and incubate at 37 °C for 1.5 h. After digestion, remove 3 μL to visualize the size distribution on a 1 % agarose gel. Completely NlaIII-digested cDNA should have a size range between 200 and 300 bp (*see* **Note 6**). Extract the reaction solution with phenol/chloroform and add 100 μL 10 *M* ammonium acetate, 3 μL glycogen, and 700 μL cold ethanol, followed by incubation at −80 °C for 1 h for ethanol precipitation. Centrifuge at 15,000 g for 40 min at 4 °C, wash the pellet once with 70 % ethanol, and dry. Resuspend the precipitate in 20 μL LoTE.
2. Prepare a 1-mL suspension of streptavidin-coated magnetic beads in a 1.5-mL microtube. Prepare two tubes of bead suspensions for each cDNA preparation. Place the tubes containing magnetic beads on a magnetic stand for 1 min and remove the supernatant with a pipet. To wash the magnetic beads, add 200 μL 1X B&W solution and suspend beads well by pipetting. Place the tube on a magnetic stand, and remove and discard the supernatant (*see* **Note 7**). Add 100 μL 2X B&W solution, 10 μL digested cDNAs (half of the cDNA from the previous step), and 90 μL distilled water to the washed magnetic beads, and suspend well. Leave the tube for 30 min at room temperature, so that the biotinylated cDNAs bind to streptavidin on the magnetic beads. Mix the bead suspension every 10–15 min. Place the tubes on the magnetic stand, and transfer supernatants, separated from beads, to new tubes. Wash the cDNA bound on the magnetic beads three times with 200 μL 1X B&W and 200 μL LoTE.
3. To ligate linkers to digested cDNAs bound to the magnetic beads, add 21 μL LoTE, 6 μL 5X T4 DNA ligase buffer, and either 1 μL Linker-1E or Linker-2E solution, respectively, to the magnetic beads. Note that Linker-1E and -2E are separately ligated to cDNAs on the magnetic beads in the two tubes. After mixing with pipets, incubate the beads suspension at 50 °C for 2 min for the dissociation of linker dimers. Keep the tubes at room temperature for 15 min. Add 2 μL T4 DNA ligase, and incubate at 16 °C for 2 h. Mix the bead suspension every 20–30 min with

a pipet. After ligation, wash beads four times with 1X B&W, and three times with LoTE (*see* **Note 8**).

4. For EcoP15I digestion of linker-cDNA on the magnetic beads, add 10 μL 10X EcoP15I digestion buffer, 2 μL 100 m*M* ATP, 83 μL sterile water, and 5 μL EcoP15I to the washed magnetic beads. Incubate beads re-suspended in reaction solution at 37 °C for 2 h with occasional mixing (every 20–30 min).

3.4. Purification of 26-bp Tags

1. Place the bead suspension on the magnetic stand, and collect the supernatant into a new tube. Add 100 μL 1X B&W to the magnetic beads, and mix. After separation on the magnetic stand, retrieve the supernatants and combine to the previously collected solution. Extract this solution, containing linker-tag fragments, with phenol/chloroform to remove the magnetic beads completely (*see* **Note 9**). Add 100 μL 10 *M* ammonium acetate, 3 μL glycogen, and 900 μL cold ethanol to the collected solution (approx 200 μL) in the tube. Keep it at −80 °C for 1 h, and centrifuge at maximum speed for 40 min at 4 °C. Wash the resulting pellet twice with 70 % ethanol, and dry by vacuum centrifuge. Dissolve precipitated linker-26 bp tag fragments in 10 μL LoTE.
2. Prepare an 8 % PAGE gel by mixing 3.5 mL 40 % acrylamide/BIS solution, 13.5 mL distilled water, 350 μL 50x TAE buffer, 175 μL 10 % ammonium persulfate, and 15 μL TEMED. Pour the solution onto the gel plate (12 cm × 12 cm, 1 mm thickness), and insert a comb (no stacking gel). Prepare running buffer, (1X TAE), and add to the upper and lower electrophoresis chambers. Then, add 2 μL 6X loading dye to 10 μL of the linker-26 bp tag solution mix, and load the sample in the well. Also, load 3 μL of a 20-bp ladder marker as molecular size marker. Run the polyacrylamide gel at 75 V for 10 min, and then at 150 V for 30 min (until the BPB dye front has migrated two-thirds down the gel).
3. After electrophoresis, remove the gel from the plate. Pour 1 mL SYBR green solution (diluted in 1X TAE buffer) on the plastic wrap, and place the gel on it. Further, disperse 1 mL SYBR green solution onto the gel and wrap the gel. After a 2-min staining period, place the gel on an ultraviolet (UV) transilluminator. Under UV light, three fragments of 46 bp, 69 bp, and 90 bp are observed (*see* **Note 10**). Only the 69-bp band (linker-26 bp tag fragments) is cut out from the gel with a spatula and transferred to a 0.5-mL microtube. Combine gel slices of both Linker1-tag and Linker2-tag in the same tube. Make holes at the top and the bottom of the tube with a needle, and place it in a 1.5-mL microtube. Centrifuge the tube at maximum speed for 2–3 min. Polyacrylamide gel pieces are collected at the bottom of the microtube. Add 300 μL LoTE to the gel pieces, and suspend. After incubation at 37 °C for 2 h, transfer the gel suspension to a Spin-X column, and centrifuge at maximum speed for 2 min. Extract the eluate with phenol/chloroform, and precipitate by adding

100 μL 10 *M* ammonium acetate, 3μL glycogen and 950 μL cold ethanol. Keep it at −80 °C for 1 h and centrifuge at 15,000 g for 40 min at 4 °C. Wash once with 70 % ethanol and dry. Dissolve the resulting pellet in 8 μL LoTE.

3.5. Ditag Formation

1. Add 1 μL 10X blunting buffer and 1 μL KOD DNA polymerase to the purified linker-26 bp tag solution (8 μL). For filling-in the ends of linker-26 bp-tag fragments, incubate at 72 °C for 2 min, and transfer onto ice immediately (*see* **Note 11**).
2. Add 30 μL LoTE and 40 μL ligation high to 10 μL blunting solution. Mix well and incubate at 16 °C for more than 4 h.

3.6. Ditag PCR Amplification

1. To optimize the concentration of template ditag for PCR amplification, prepare 10X and 20X diluted ligation mixtures (from the previous step) with LoTE (each 10 μL), and check for PCR efficiency.
2. Prepare ditag PCR reaction mixture, containing 5 μL GeneAmp 10X PCR Gold Buffer, 5 μL 2 m*M* dNTP, 6 μL 25 m*M* $MgCl_2$, 0.2 μL biotinylated Ditag1E primer, 0.2 μL biotinylated Ditag2E primer, 32.34 μL distilled water, 1μL diluted template solution, and 0.26 μL AmpliTaq Gold.
3. PCR cycle: 94 °C for 12min, then 27–29 cycles each at 94 °C for 40 s, and 60 °C for 40 s.
4. Prepare 8 % polyacrylamide gel as described under **Subheading 3.4.**, and load each 6-μL ditag PCR product to the gel. After SYBR green staining, bands of PCR products are visualized under UV light. Determine the optimal dilution of the ditag template (*see* **Note 12**), judging from intensity of 96-bp amplified ditag, 45-bp amplified linker-dimers, and fragments of other sizes (**Fig. 1B**).
5. Once conditions for PCR amplification are optimized, run 40 PCR reactions (40 tubes). Add a blank control (1 μL distilled water is used instead of template; *see* **Notes 13** and **14**).

3.7. Purification of Ditag PCR Products and Second NlaIII Digestion

1. Collect all the PCR solutions (40 reactions, 2 mL) in a 15-mL tube, and add 10 mL binding buffer (PB buffer). Prepare eight Qiaquick spin columns from the Qiaquick PCR purification kit, and transfer 750 μL of a mixture between PCR solution and PB buffer to each column. Centrifuge at 10,000 g for 1 min, and discard flow-through. Add another 750 μL of residual mixture to each column. After centrifugation at 10,000 g for 1 min, transfer columns to new tubes, and add 750 μL washing buffer (PE buffer, ethanol added) to each column. Centrifuge at 10,000 g for 1 min, and discard flow-through. For completely drying the columns, centrifuge at maximum speed for 2 min. Transfer columns to new 1.5-mL microtubes, and add 30 μL LoTE to each column for elution. Centrifuge at 10,000 g for 1 min, and collect eluate.

2. Collect purified ditag PCR products (eluate from the column) in two tubes (approx 120 μL in each tube). Add 15 μL NlaIII digestion buffer (NEBuffer 4), 1.5 μL 100X BSA, and 12 μL NlaIII, and incubate at 37 °C for 1 h.
3. Prepare 500 μL streptavidin-coated magnetic beads, and wash them once with 200 μL 1X B&W solution. Remove 3 μL digestion solution, and run on an 8 % PAGE to check digestion with NlaIII. If most of the 96-bp band is digested, and the 52-bp band is clearly visible, all of the digestion solution is transferred to the washed streptavidin magnetic beads and mixed well by pipeting (*see* **Note 15**). Leave at room temperature for 20 min. Place the tube on the magnetic stand and transfer supernatant to a new tube. Add 100 μL 1X B&W to the magnetic beads, and mix. After separation on the magnetic stand, collect supernatants and combine with the previously collected solution. Extract with phenol/chloroform (*see* **Note 15**). Add 100 μL 10 *M* ammonium acetate, 3 μL glycogen, and 900 μL cold ethanol to the collected solution. Keep it at −80 °C for 1 h, and centrifuge at maximum speed for 40 min at 4 °C. Wash the pellet once with 70 % ethanol, dry, and dissolve in 10 μL LoTE.
4. Prepare a 12 % polyacrylamide gel (12 cm × 12 cm) by mixing 5.25 mL 40 % acrylamide/BIS solution, 11.75 mL distilled water, 350 μL 50X TAE buffer, 175 μL 10 % ammonium persulfate, and 15 μL TEMED. Load the sample in 2–4 lanes (*see* **Note 16**), and run the gel as described under **Subheading 3.4**. Stain the gel with SYBR green and visualize the bands on the UV transilluminator. Cut out only the 52-bp bands (ditag) from the gel and transfer to a 0.5-mL tube. Elute DNA from the polyacrylamide gel using the procedure described in **Subheading 3.4.** To completely eliminate biotinylated fragments, treat the eluted solution (300 μL) with streptavidin magnetic beads (washed once with 1X B&W before use) once again, and extract the collected supernatant with phenol/chloroform (*see* **Note 17**). Ethanol precipitate by adding 100 μL 10 *M* ammonium acetate, 3 μL glycogen, and 900 μL cold ethanol to the supernatant, and mix. Keep it at −80 °C for 1 h, and centrifuge at 15,000 *g* for 40 min at 4 °C. Wash the resulting pellet twice with 70 % ethanol, dry, and dissolve in 6 μL LoTE.

3.8. Concatemer Formation and Purification

1. Add 6 μL ligation high mixture to the purified 52-bp ditag solution, and incubate at 16 °C for more than 3 h.
2. Prepare a 6 % polyacrylamide gel (12 cm × 12 cm) by mixing 2.6 mL 40 % acrylamide/BIS solution, 14.4 mL distilled water, 350 μL 50X TAE buffer, 175 μL 10 % ammonium persulfate, and 15 μL TEMED. After 1 h of ligation reaction, remove 1 μL for an electrophoresis in a 6 % polyacrylamide gel. If the concatemer is properly formed, smear or ladder DNA is observed above 100 bp.
3. Add 10 μL LoTE and 300 μL buffer ERC from the MinElute Reaction Cleanup Kit to the ligation reaction. Transfer the mixture to a MinElute Spin Column, and centrifuge at 10,000 g for 1 min. Discard flow-through and add 750 μL washing

buffer (PE buffer containing 80 % ethanol). After centrifugation at 10,000 g for 1 min, spin for an additional 1 min at maximum speed. Place the column in a new 1.5-mL microtube and add 10 μL LoTE. Leave it for 1 min at room temperature and centrifuge at 10,000 *g* for 1 min. Then, add 1.4 μL NlaIII digestion buffer (NEBuffer 4) and 1.4 μL 10X BSA to the eluate. Start digestion with 1 μL NlaIII and incubate at 37 °C for partial digestion of concatemers. After 30–60 s of incubation (*see* **Note 18**), place the tube to 75 °C immediately, and incubate for 20 min. Afterward, transfer on ice. This NlaIII treatment linearizes circular concatemers as explained in the robust-LongSAGE protocol ***(11)***.

4. Prepare a 6 % polyacrylamide gel (16 cm × 16 cm) by mixing 5.2 mL 40 % acrylamide/BIS solution, 28.8 mL distilled water, 700 μL 50 × TAE buffer, 350 μL 10 % ammonium persulfate, and 30 μL TEMED. Load the partially digested concatemers, and both the 20-bp and 100-bp ladder markers in adjacent lanes. Run the gel at 150 V for approx. 2 h. Visualize DNA within the SYBR green-stained gel on a UV transilluminator. Cut out the concatemers in the size range between 500 and 1000 bp, and between 350 and 500 bp separately (*see* **Note 19**). Transfer each gel to a 0.5-mL tube separately, and elute DNA as described above **(Subheading 3.4.)**. After phenol/chloroform extraction and ethanol precipitation, dissolve purified concatemers in 6 μL LoTE.

3.9. Vector Cloning and Transformation

1. For digestion of the pGEM-3Z vector, mix 1 μL plasmid, 2 μL SphI digestion buffer (NEBuffer 2), 15 μL distilled water, and 2 μL SphI, and incubate at 37 °C for 1 h. Add 100 μL PB buffer of the Qiaquick PCR purification kit to the digestion solution, and purify the plasmid DNA with a Qiaquick spin column, as described under **Subheading 3.7.** After washing with PE buffer, elute the plasmid from the column with 30 μL LoTE. Add 5 μL 10X alkaline phosphatase buffer and 1 μL calf intestine alkaline phosphatase, and incubate at 50 °C for 30 min. Purify the reaction with a Qiaquick spin column from the Qiaquick PCR purification kit again. Adjust the concentration of the purified plasmid to 5 ng/μL in LoTE.
2. To 6 μL of the purified concatemer solution, add 1 μL SphI-digested pGEM-3Z and 7 μL ligation high solution, and incubate at 16 °C for 4 h. Add 300 μL buffer ERC and purify using a MinElute Spin Column from the MinElute Reaction Cleanup kit as described under **Subheading 3.8.** Elute the plasmid from the column with 10 μL LoTE.
3. Place electrocompetent *E. coli* cells (ElectroMAX DH10B) on ice. In a microtube on ice, mix 20 μL competent cell suspension and 1 μL plasmid solution. Transfer the mixture to a chilled 1-mm-distance electroporation cuvet, and electroporate the cells using the following conditions: 2.0 kV, 200Ω., and 25 μF. Recover electroporated cells in 1 mL SOC medium, and incubate at 37 °C for 1 h with shaking.
4. Plate transformed cells on LB medium containing ampicillin, IPTG, and X-gal. Incubate the plates at 37 °C overnight.

5. Prepare a PCR mixture by mixing 2 μL PCR buffer, 1.8 μL dNTP, 1 μL M13F primer (20 pmol/μL), 1 μL M13RV primer (20 pmol/μL), and 4.2 μL distilled water (per PCR tube). Pick a white colony by a toothpick, suspend in 10 μL distilled water in a PCR tube, and add 10 μL PCR mixture.
6. Amplify plasmid inserts by the following PCR program: 94 °C for 20 s, 30 cycles each of 55 °C for 40 s and 72 °C for 2.5 min, and 72 °C for 10 min. Run PCR products on a 1 % TAE agarose gel to estimate the average size of the cloned concatemers (*see* **Note 20**).
7. Purify PCR products and sequence them directly. Analyze sequence data with the SAGE2000 program for 22-bp tags (excluding the common CATG at the 5′-end of tags).

Acknowledgments

This work was supported by "Program for Promotion of Basic Research Activities for Innovative Biosciences" (Japan), Grant-in-Aid for Young Scientist (A) 1868801 and by "Iwate University 21st, Century COE Program". Research of P. Winter and G. Kahl is supported by the European Union (contract FOOD CT-2004–506223) and IAEA (CRP 302–041-GFR-8148). Work of D. H. K. and M. R. was supported by Deutsche Forschungsgemeinschaft grant Kr 1293/4.

4. Notes

1. Streptavidin-coated magnetic beads supplied from other companies can also be used. However, biotin-binding capacity of the beads may differ from that of Promega, so that volume of beads suspension for use should be optimized. For instance, DynaBeads (Dynal Biotech, Norway), 100 μL suspension, have the same binding capacity as 1 mL of Promega's beads.
2. M. Reuter, D. H. Krüger, and co-workers are able to overexpress and purify EcoP15I enzyme, which has both a higher catalytic activity and stability as compared to the enzyme available from New England Biolabs, and is therefore especially suited for SuperSAGE ***(12)***.
3. Never use the buffer and other solutions attached to EcoP15I enzyme of New England Biolabs. Only the buffer system described herein is acceptable for obtaining sufficient amounts of linker-26 bp tags after EcoP15I digestion.
4. If most of the synthesized cDNA is less than 500 bp, then mRNA might be degraded. It is wise to prepare a new mRNA sample.
5. Phenol/chloroform extraction in other steps also follows this procedure.
6. If you cannot see any down shift of cDNA sizes after NlaIII digestion, you should not move on. In such a case, first check the activity of NlaIII, and then confirm whether double-stranded cDNA is properly synthesized.
7. Magnetic beads washing in other steps also follows this procedure.
8. During washing with 1X B&W or LoTE, tubes should repeatedly be replaced by new ones. This washing step is important for eliminating unligated linkers

and linker dimers. When substantial amounts of these molecules remained, they compete with ditags during PCR reaction and yield of ditag PCR products is remarkably reduced.

9. Phenol/chloroform extraction at this step eliminates contaminated magnetic beads from the solution.
10. Unligated linker is 46 bp, and linker dimer 90 bp in length. Mobility of each fragment is delayed, because the bulky fluorescein isothiocyanate (FITC) is attached to the end. Sometimes, the linker dimer is not visible. If the intensity of the 69-bp band (linker-26 bp tag) is faint, while the other two bands are clearly visible, EcoP15I digestion is inefficient, or the amount of cDNA on the magnetic beads is too low. If no band is visible, linkers might not be ligated to cDNA. Any contamination of linker-26 bp tags by unligated linkers or linker dimers should be avoided when cutting out the fragment from the gel.
11. In place of KOD DNA polymerase, other DNA polymerases (T4 DNA polymerase or Klenow fragment) can be used for blunting. When other DNA polymerases are used, purification of blunt-ended linker-26 bp tag fragments is required before proceeding to the ligation reaction.
12. When preparing the ditag PCR mixture, care should be taken so that no contamination of previously amplified PCR products occurs. Use separate pipets and solutions, including water, from those used in the experiments after ditag PCR. Also, use separate labware and gloves. If ditag amplification is observed in the blank control lane, all the PCR solutions and tips should be renewed, and pipets should be irradiated with UV light in the clean bench for 15 min to eliminate contaminating DNA.
13. If you see intensive amplification of fragments of sizes other than 96 bp even in 20X diluted template, you may dilute further, provided that the 96-bp fragment is still well amplified.
14. No amplification of ditags should be observed in the negative control as shown in **Fig. 1B**. If cross-contamination is observed, discard all the solutions and restart from the ditag PCR amplification step **(Subheading 3.6.)**.
15. Most of the biotinylated fragments of the linker region (primers) are trapped on the magnetic beads ***(13)***, and the 52-bp fragment is clearly separated from other fragments on a polyacrylamide gel.
16. Loading too much DNA to the well reduces resolution of PAGE. However, when separated in too many lanes (more than five lanes), the volume of the resulting gel slices is increased and recovery of DNA from the gel may be suboptimal.
17. To obtain good quality of concatemers, complete elimination of biotinylated fragments is essential.
18. The time for partial digestion with NlaIII depends on the amount of concatemer molecules. If most concatemers are less than 500 bp after partial digestion, or if DNA is not visible in the polyacrylamide gel, more ditag PCR products

(corresponding to 50–60 PCR tubes) should be subjected to NlaIII digestion and concatemer formation.

19. Avoid contamination of the larger concatemers with small fragments when you cut out the fragments from the gel. Small fragments (<200 bp) are more likely to be cloned into the vector, even if their amount is low.
20. If many small fragments are amplified (>50 % of clones), return to the ditag PCR step **(Subheading 3.6)** and increase the number of tubes for ditag PCR (50–60 tubes).

References

1. Tan, P. K., Downey, T. J., Spitznagel, E. L., et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucl. Acids Res.* **31,** 5676–5684.
2. Suárez-Fariñas, M., Noggle, S., Heke, M., Hemmati-Brivanlou, A., and Magnasco, M. O. (2005) Comparing independent microarray studies: the case of human embryonic stem cells. *BMC Genomics* **6,** 99–109.
3. Ding, C. and Cantor, C. (2003) High-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc. Natl. Acad. Sci. USA* **100,** 3059–3064.
4. Brenner, S., Johnson, M., Bridgham, J., et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotech.* **18,** 630–634.
5. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270,** 484–487.
6. Matsumura, H., Reich, S., Ito, A., et al. (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. USA* **100,** 15718–15723.
7. Matsumura, H., Reich, S., Reuter, M., et al. (2004) SuperSAGE: A transcriptome tool for eukaryotic organisms, in *SAGE: Current Technologies and Applications* (Wang, S. M., ed.). Horizon Scientific Press, Norwich: pp. 77–90.
8. Matsumura, H., Ito, A., Saitoh, H., et al. (2005) SuperSAGE. *Cell. Microbiol.* **7,** 11–18.
9. Kim, D-H., Behlke, M. A., Rose S. D., Chang, M-S., Choi, S. and Rossi, J. J. (2005) Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy. *Nat. Biotech.* **23,** 222–226.
10. Siolas, D., Lerner, C., Burchard, J., et al. (2005) Synthetic shRNAs as potent RNAi triggers. *Nat. Biotech.* **23,** 227–231.
11. Gowda, M., Jantasuriyarat, C., Dean, R. A., and Wang, G.L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol.* **134,** 890–897.
12. Moncke-Buchner, E., Mackeldanz, P., Kruger, D. H, and Reuter, M. (2004) Overexpression and affinity chromatography purification of the Type III

restriction endonuclease EcoP15I for use in transcriptome analysis. *J. Biotechnol.* **114,** 99–106.

13. Powell, J. (1998) Enhanced concatemer cloning—a modification to the SAGE (Serial Analysis of Gene Expression) technique. *Nucl. Acids Res.* **26,** 3445–3446.

5

Low-Cost-Medium Throughput Sanger Dideoxy Sequencing

Kåre Lehmann Nielsen

Summary

Serial analysis of gene expression (SAGE) requires the sequencing of DNA. The principal cost of SAGE is largely determined by the cost of sequencing. Therefore, it is important to have access to a robust and affordable sequencing system. Here, we describe such a system based on the sequencing of amplified inserts of concatemer-containing clones.

Key Words: DNA sequencing; medium throughput; PCR products; concatemers.

1. Introduction

Serial analysis of gene expression (SAGE) and LongSAGE are digital gene expression profiling methods that relies on the counting of DNA sequences ***(1,2)***. For almost 30 yr, DNA sequencing has almost exclusively been conducted in accordance with the dideoxy chain-terminating method described by Sanger ***(3)***. Lowering of costs and enormous progress in reliability, speed, and throughput have been achieved in the last decade, mostly as a result of the technology developments fueled by the human genome project ***(4,5)***. Today, 3×96 DNA sequences of 750 bp (10–15 LongSAGE ditags) are routinely obtained in an ordinary molecular biology laboratory equipped with a 96-channel capillary DNA sequencer by a single technician per day. However, in order to capitalize on the remarkable throughput of the DNA sequencer, a robust, low-cost sequence template preparation pipeline with minimal handling time is required.

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

PCR *(6,7)* products constitute excellent DNA sequencing templates for several reasons: the DNA fragments are relatively small, they are virtually free of contaminating DNA, and uniform DNA concentrations can be obtained from a wide size range of clone inserts, which is important for the processing of multiple samples. Furthermore, they are relatively cheap, and 96 or even 384 reactions can readily be prepared simultaneously.

2. Materials

2.1. Picking and Propagation of Plasmid-Containing Bacteria

1. Standard toothpicks, autoclaved in beakers.
2. Sterile 96-well microtiter plate (round well) from Greiner-BioOne (Frickenhausen, Germany).
3. Autoclaved low-salt Luria-Bertani (LB) liquid media, pH 7.2: 10 g/L tryptone, 5 g/L yeast extract, and 5 g/L NaCl, (Bie og Berntsen, Rødovre, Denmark). Store in the dark at room temperature.
4. LB plates containing 1.2 % agar (Bie og Berntsen) and appropriate antibiotics with bacterial colonies containing plasmids for sequencing. Store for a maximum of 7 d at 4 °C.
5. 8- or 12-multichannel pipet from Biohit (Helsinki, Finland).

2.2. Plasmid Mini Prep in 96-Well Plates

1. GTE solution: 50 m*M* glucose, 25 m*M* Tris-HCl, pH 8.0, and 10 m*M* ethylenediamine tetraacetic acid (EDTA). Autoclave and store at room temperature.
2. Lysis solution: 0.2 *M* NaOH and 1 % (w/v) sodium dodecyl sulfate (SDS). Stock solutions of 0.4 M NaOH and 2 % (w/v) SDS can be made and stored at room temperature. Mix 1:1 immediately before use.
3. Neutralization solution: prepare a 5 *M* potassium acetate, pH 4.8, by slowly adding KOH pellets to 29.5 mL of glacial acetic acid on ice until the pH is 4.8 (*see* **Note 1**). Add water (*see* **Note 2**) to 100 mL. Store at room temperature.
4. 2-propanol (Sigma-Aldrich, St. Louis, MO).
5. 70 % EtOH (De danske spritfabrikker, Aalborg, Denmark)
6. Laboratory tissues.

2.3. Amplification of Plasmid Inserts by PCR

1. Taq polymerase 1 U/μL from Fermentas (Burlington, Ontario, Canada). Store at −20 °C.
2. Taq reaction buffer (10X): 100 m*M* Tris-HCl (pH 8.8), 500 m*M* KCl, 8 % Nonidet P40.
3. Deoxynucleotides, dATP, dTTP, dGTP, and dCTP (100 m*M* stock solutions) from Fermentas. Prepare a mix of all four (final concentration 25 m*M* of each). Store at −20 °C.

4. 25 m*M* $MgCl_2$. Chemicals for PCR are conveniently stored together at −20 °C.
5. Two vector-specific primers (10 μ*M*). Store at −20 °C.
6. 96-well PCR microtiter plate.

2.4. TAE-Agarose Gel Electrophoresis

1. TAE running and gel buffer (50X): 2 *M* Tris, 1 *M* acetic acid, and 50 m*M* EDTA, pH 7.6–7.8. Store at room temperature.
2. SeaKem GTG agarose (Cambrex, East Rutherford, NJ). Stored dry at room temperature.
3. Gel casting and running apparatus Mini-Sub Cell GT from Biorad (Hercules, CA).
4. EtBr solution: 10 g/L ethidium bromide (Sigma-Aldrich) in water. Store at 4 °C.
5. TAE loading buffer (5X): 0.2 *M* Tris, 0.1 *M* acetic acid, 5 m*M* EDTA in 50 % glycerol. Store at 4 °C.
6. DNA size markers: 1 kb GeneRuler™ (Fermentas). Store at 4 °C.

2.5. ExoSAP Digestion of PCR Product

1. Shrimp alkaline phosphatase (1 U/μl) from Fermentas.
2. SAP buffer (10X): 0.1 *M* Tris-HCl, pH 7.5, 0.1 *M* $MgCl_2$, and 1 mg/mL bovine serum albumin (Fermentas). Store at −20 °C or 4 °C.
3. Exonuclease I (20 U/μL) from Fermentas. Store at −20 °C.

2.6. DNA Cycle Sequencing Protocol

1. 40 μ*M* sequencing primer (*see* **Note 3**).
2. DYEnamic ET terminator mix (GE-healthcare, Chalfont St. Giles, UK) (*see* **Note 4**)
3. 96-well PCR microtiter plate.

2.7. Purification of the Sequencing Reaction by Gel Filtration

1. Millipore (MAHV-N45) filter plate (*see* **Note 5**)
2. Sephadex G-50 fine dry powder (GE-Healtcare). Store at room temperature.
3. Filter plate loading device (GE-Healtcare).
4. 96-well PCR microtiter plate.
5. Alignment rings (GE-Healtcare)
6. Homemade alignment ring.
7. Centrifuge for microtiter plates.

3. Methods

Most DNA sequencing failure is caused by insufficient quality or inappropriate amount of template DNA. Contrary to what most people think, failure in capillary-based DNA sequencing is more often caused by too much template DNA rather than too little. In fact, a single microliter of a reasonable PCR product (50–100 ng) usually provides the best results. Therefore, when

designing the present reaction setup, more value was attributed to consistency, reliability, and minimal handling than to maximum yield. Using this protocol, we usually process four 96-well plates in parallel and routinely obtain a satisfactory DNA sequence from 74 % of clones picked.

3.1. Picking and Propagation of Plasmid-Containing Bacteria

1. Add 150 µL LB agar containing 100 µg/µL Zeocin (*see* **Note 6**).
2. From agar plates containing bacterial colonies, pick all colonies into the low-salt LB-Zeocin-containing microtiter plate, one colony in each well (*see* **Note 7**).
3. Seal the plate and leave in a 37 °C incubator overnight. Alternatively, incubate in the dark at room temperature for 2 d (or a weekend) (*see* **Note 8**).

3.2. Plasmid Mini Prep in 96-Well Plates

1. Spin the culture plates for 5 min at 900*g* in a microtiter plate-compatible centrifuge.
2. Discard supernatant by inverting the plate (*see* **Note 9**). Wipe off the plate with a tissue while still inverted.
3. Using a multichannel pipet, add 50 µL of GTE solution, seal the plate, and vortex for 20 s. Spin the plate for 10 s at 900*g* to collect the reaction at the bottom of the well (*see* **Note 10**). Remove the seal (*see* **Note 11**).
4. To lyse the bacteria, add 100 µL lysis solution to all wells. Mixing is not necessary at this step.
5. To precipitate most of the proteins, the genomic DNA and cell debris, add 50 µL of neutralization solution to all wells. A big white precipitate will form. Spin for 15 min at 900*g* and transfer 100 µL of the clear supernatant to a new plate (*see* **Note 12**).
6. To precipitate the polynucleic acids from the 100 µL of supernatant, add 75 µL 2-propanol, and incubate at −20 °C for 30 min (*see* **Note 13**).
7. Centrifuge at 900*g* for 15 min and discard the supernatant by inverting the plate. Wipe off the inverted plate. The precipitate is now visible as a faint smear all over the bottom of the well.
8. Carefully add 150 µL of 70 % ethanol slowly down the side of the wells. Discard the supernatant by inverting the plate and add another 150 µL of 70 % ethanol (*see* **Note 14**). Discard the supernatant by inverting the plate and wipe off the inverted plate carefully with a tissue. Leave the plate on the bench top for 10 min to evaporate residual ethanol.
9. To dissolve the DNA, add 100 µL of water to each well, seal the plate and vortex for 20 s. Leave it at room temperature for 15 min and vortex for another 20 s. Spin briefly at 900*g* for 10 s to collect the reaction in the bottom of the wells. The plasmid mini preps can be used immediately or stored at −20 °C.

3.3. Amplification of Plasmid Inserts by PCR

1. Set up the thermocycler with the following profile: after initial denaturation at 94 °C for 3 min, 25 cycles of 94 °C for 30 s, 53 °C for 30 s, and 72 °C for 105 s should be followed by 7 min at 72 °C.
2. To each well of a 96-well PCR plate, add 10 μL of water and then 2.5 μL of the plasmid mini preps from under **Subheading 3.2.** (*see* **Note 15**).
3. Prepare a mastermix for 96 PCR reactions by combining on ice in the following order: 570 μL water, 250 μL Taq buffer (10X), 25 μL dNTP mix, 50 μL of each primer, 300 μL of $MgCl_2$, and 30 μL of Taq polymerase. Mix gently (*see* **Note 16**). Add 12.5 μL to each well and seal the plate (*see* **Note 17**).
4. Allow the thermocycler to preheat to 75 °C before inserting the plate into the thermocycler.
5. Analyze 2 μL on a TAE agarose gel (*see* **Note 18**).

3.4. TAE-Agarose Gel Electrophoresis

1. Mix 0.5 g agarose (for a 1 % gel) with 49 mL of water and 1 mLof 50X TAE. Melt agarose in microwave (*see* **note 19**) and add 2 μL of EtBr.
2. Seal a gel casting form with tape (*see* **Note 20**), pour the melted agarose into the form, and insert a comb. Leave on the bench top until hardened.
3. Remove comb and insert in horizontal gel apparatus and submerge in 1X TAE buffer (*see* **Note 21**).
4. Mix the sample with one-fifth volume of 5X loading buffer and load into wells. Include a suitable DNA size marker (*see* **Fig. 1**).
5. Electrophorese at 10 V/cm until the bromphenol blue is about half the distance of the gel (*see* **Note 22**).
6. Visualize DNA bands by exposure to ultraviolet light.

Figure 2 shows a typical sequencing electropherogram from templates prepared as described.

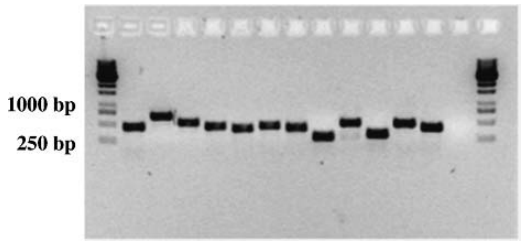


Fig. 1. PCR products suitable for DNA sequencing. Lanes 1 and 15, 1-kb DNA size marker. Lanes 2–13, 2- μL PCR product from long-serial analysis of gene expression concatemer containing pZErO-1 plasmid preparations. Lane 14, negative control (no template).

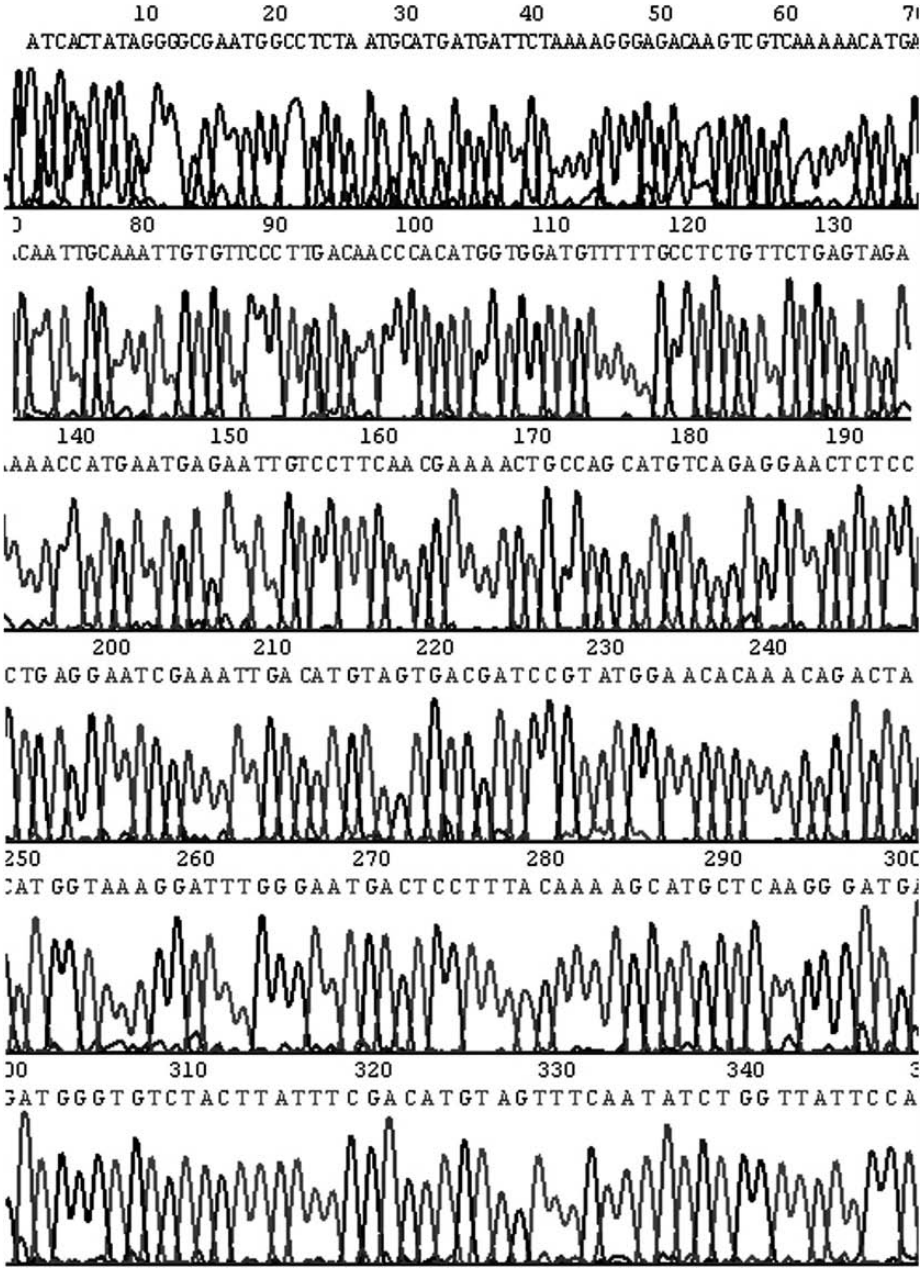


Fig. 2. Typical sequence sequencing electropherogram from templates prepared as described.

3.5. ExoSAP Digestion of PCR Product

1. Prepare a digestion mix for 96 samples by combining on ice 202 µL of water, 268 SAP buffer (10X), and 32 µL of shrimp alkaline phosphatase and 32 µL of exonuclease I.
2. To digest excess primers and triphosphate nucleotides from the PCR reactions, add 5 µL of digestion mix to each well. Incubate at 37 °C for 1 h followed by 80 °C for 20 min (*see* **Note 23**).
3. Spin 10 s at 900*g* to collect any condensation into the bottom of the wells.

3.6. DNA Cycle Sequencing Protocol

1. Set up the thermocycler with the following profile: 25 cycles of 94 °C for 20 s, 57 °C for 15 s, and 60 °C for 1 min followed by 5 min at 60 °C (*see* **Note 24**).
2. Assemble the sequencing reaction by adding 6 µL of water and 1 µL of PCR product from under **Subheading 3.5.** to each well of a microtiter PCR plate (*see* **Note 25**).
3. In a separate tube, add 51 µL of a single sequencing primer to 510 µL of DYEnamic ET terminator mix. Add 5 µL to each well of the microtiter plate (*see* **Note 26**).
4. Insert into the thermocycler and run the program in **step 1**.

3.7. Purification of the Sequencing Reaction by Gel Filtration

Salts and unincorporated dye terminators must be removed prior to injection into the capillary sequencer. Excess salts will disturb the electrokinetic injection process and excess label will result in "dye blobs" in the electropherogram, decreasing read length (*see* **Note 27**).

1. Fill a filter plate with dry Sephadex G-50 fine using a special loading device (GE-Healthcare) to ensure equal amounts of G50 is loaded into each well (*see* **Note 28**). Add 300 µL of water to each well, put a lid on the plate, place the plate in a plastic bag, and let the matrix swell at room temperature for at least 1 h (or overnight at 4 °C) (*see* **Note 29**).
2. Put the swelled filter plate (with the lid on) on top of a regular 96-well microtiter plate, using the alignment ring to fix them together. Spin the assembly for 5 min at 900*g*. Discard the flow-through from the lower plate. Add another 150 µL of water to each well and put the lid on. Place it on top of the microtiter plate using the alignment ring and spin for another 5 minutes at 900*g*. Discard the flow-through.
3. Place the filterplate on top of a PCR plate that is compatible with the MegaBace loading dock using an appropriate alignment ring (*see* **Note 30**). Immediately, add the sequencing reactions to the wells containing the G-50 matrix. Be careful to add them in the center of the matrix in each well without disturbing the matrix (*see* **Note 31**). Put the lid on the filter plate, tape the assembly together, and spin at 900*g* for 5 min.
4. Disassemble the plates and add 10 µL of water to the approx 8 µL of purified sequencing reaction in each well of the lower PCR plate (*see* **Note 32**). Spin the

plate briefly for 10 s to remove any bubbles; the plate is now ready for loading into the MegaBace (*see* **Note 33**).

5. Discard the used G-50, rinse the filter plate thoroughly with water, leave to dry on the bench top overnight, and reuse.

4. Notes

1. Be careful when preparing the neutralization solution. The acid–base reaction liberates a substantial amount of heat, which will cause the solution to boil if not prepared slowly on ice.
2. Unless stated otherwise, all solutions should be prepared in water that has a resistivity of 18 MΩ/cm and an organic content of less than five parts per billion. However, for small volume solutions up to 10 mL, molecular biology-grade water from AppliChem can be used to ensure reproducibility even when the local water systems fail.
3. Not all primers work equally well as sequencing primers, and some protocols recommend the use of a primer different from the two used to amplify the product. We have observed no problems, however, with using the same primers as for the amplification. We use the following primers, which are found in many cloning vectors: M13-21 (5′-GTAAAACGACGGCCAG-3′), M13 rev (5′-GCAGGAAACAGCTATGAC-3′), T7 (5′-TAATACGACTCACTATAGGG-3′), and T7rev (5′-GCTAGTTATTGCTCAGCGG-3′).
4. We are using a MegaBace 1000 DNA sequencer and thus are using the GE-Healthcare sequencing chemistry.
5. The filter plates are quite expensive. However, if rinsed in water and left to dry, they can be recycled at least 25 times without significant carryover of DNA sequence signal.
6. The pZErO-1 vector used for cloning of SAGE concatemers confers resistance to Zeocin. If others vectors are used, adjust the antibiotics accordingly.
7. To keep track of the picking process, leave the toothpick in the well until an entire row in the plate is completed.
8. Do not shake the plate. The seal will leak and the cultures will contaminate each other. A sufficient amount of bacteria is obtained without shaking.
9. Inversion of the plate must take place immediately after centrifugation ends to ensure that the bacterial pellets stay in the microtiter plate.
10. The short spin is merely to collect droplets deposited on the seal into the bottom of the wells. The bacteria should not pellet in this step.
11. Press the plate firmly against the bench top while removing the plate seal in order to avoid spills and cross contaminations.
12. Try to avoid getting any precipitate into the new plate. A small amount of precipitate, however, does not influence the result.
13. If using cold 2-propanol (−20 °C), no incubation time is necessary.

14. It is possible to wash with ethanol without disturbing the precipitate. Alternatively, the plate can be spun at 2600*g* for 10 min between each wash.
15. Do not leave the droplet at the side of the well, but pipet directly into to the liquid already in the wells. It is important to have a reasonable amount of liquid (10 μL) already in the wells to consistently and reliably transfer 2.5 μL using a multichannel pipet.
16. Do not vortex the solution; intracellular enzymes are easily inactivated by oxidation of cysteine residues. Instead, mix gently by pipetting up and down three times with a 1-mL pipet, set at 500 μL, avoiding bubbles.
17. Complete sealing of the plate is, of course, necessary in order to avoid loss of the samples due to evaporation. We find silicone mats from Greiner very convenient for this purpose. We recycle them by submerging them in 1 *M* NaOH for at least 1 h and subsequently rinsing them with water and 70 % ethanol.
18. We routinely run 12 samples from a plate on a 1 % TAE-agarose gel for monitoring of sequence template quality. When troubleshooting or modifying protocols, we run all 96 samples.
19. Use a 100-mL bottle for 50 mL gel to avoid boiling over. Put the lid on loosely to minimize loss by evaporation.
20. Not all types of tape will stay sealed. Autoclave tape works fine and is often at hand but is the most expensive. Functional, cheaper alternatives can be found.
21. Add just enough buffer to submerge the gel. Too much buffer will only increase electrophoresis time and buffer consumption.
22. The bromphenol blue dye migrates as approx 400 bp in a 1 % TAE-agarose gel.
23. The reaction is most conveniently carried out in a thermocycler.
24. The optimal annealing temperature (T_{ann}) of the cycle sequencing reaction depends on the sequence primer used. Optimal T_{ann} tends to be higher than predicted by melting temperature calculations made by PCR primer design software. However, the sequence reaction does not seem to be very sensitive to small changes in T_{ann}, and we are using the same temperature for the four primers listed in materials.
25. It is important not to add too much PCR template to the sequence reaction. Typically, a sequence reaction will yield acceptable data in the range of 0.2–3 μL of PCR product.
26. In fact, the volume can be scaled down to 8 μL. But inaccuracy in pipetting increases and more reactions fail.
27. There are several methods of desalting sequencing products: ethanol precipitation, binding of biotin-labeled sequencing products to paramagnetic beads, or even simple dilution. However, we have found that Sephadex G-50 gel filtration works well and consistently, and by recycling the filter plates, it is not expensive.
28. It is very important that the same amount of G50 is loaded in each well and from plate to plate. The washing volumes are finely tuned to match the volume of the gel filtration matrix. If too large a matrix volume is used, unincorporated label will contaminate the sequencing products and the yield of sequencing products

will be lowered. If too small a matrix volume is used, the sequencing product will be eluted in the washing steps.

29. We usually load 25 plates at a time with dry powder and leave them at room temperature in a plastic bag. Plates that have been swelled should be used the same day. If plates have been swelled overnight in the cold, they must be incubated at room temperature for at least 1 h before use.
30. We are using a homemade alignment ring cut from an acryl plate to ensure that the filter plate is firmly in place on top of the PCR plate.
31. Reliable transfer is most easily achieved by using an eight-channel pipet set at 10–12 μL.
32. Occasionally, we get an uneven amount of eluted sample volume across the plate. It can often be rectified by rotating the plate assembly 180° and spinning for another 5 min at 900*g*.
33. We use 40 s injection at 3 kV for this type of sample.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270,** 484–487.
2. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20,** 508–512.
3. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
4. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
5. Lander, E. S., Linton, L. M., Birren,B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
6. Saiki, R. K., Scharf, S., Faloona, F., et al. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230,** 1350–1354.
7. Mullis, K. B. and Faloona, F. A. (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155,** 335–350.

6

DeepSAGE

Higher Sensitivity and Multiplexing of Samples Using a Simpler Experimental Protocol

Kåre Lehmann Nielsen

Summary

Combining serial analysis of gene expression (SAGE) with pyrophosphatase-based ultra-high-throughput DNA sequencing provides increased sensitivity and cost-effective gene expression profiling. The combined techniques obviate the formation and cloning of concatemers and the tedious picking and preparation of sequence templates from bacterial clones that are necessary with SAGE alone. Furthermore, multiplexing of samples or replicates of analysis is included in the experimental design.

Key Words: Pyrophosphatase DNA sequencing; ditag sequencing; high throughput; deep sampling.

1. Introduction

Serial analysis of gene expression (SAGE) is a high-throughput method for global gene expression analysis ***(1)***. SAGE is based on two principles: that a short nucleotide sequence (tag) from a unique position contains sufficient information to uniquely identify a transcript, and that the tags, in contrast to the full-length transcripts, can be amplified without altering relative abundances. Tags are isolated, ligated together, cloned, and sequenced. In a typical sequence run of 96 samples, approx 1500 tags, and therefore, mRNAs, are detected. Because of the cost of sequencing, a SAGE study typically encompasses 50,000 tags and provides detailed knowledge of the 2000 most highly expressed genes

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

in the tissue analyzed. In practice, however, it has proven difficult to achieve enough clones of the appropriate insert length ***(2)*** to facilitate efficient detection.

Recently, Margulies et al. ***(3)*** presented a new ultra-high-throughput DNA sequencing technology that, instead of isolating DNA sequence templates from bacteria, used paramagnetic beads and PCR for the binding and clonal amplification of DNA fragments. Routinely, more than 200,000 DNA sequences of about 120 nt of which approx 80 nt is of high quality, are obtained. The fact that a LongSAGE ditag is about 40 nt long called for an integration of the two technologies based on sequencing of SAGE ditags instead of concatemers. The resulting method, DeepSAGE, is an experimentally simple method of tag-based transcript detection that is similar to LongSAGE ***(4)*** but which, in conjunction with emulsion-based amplification and pyrophosphate-based ultra-high-throughput DNA sequencing ***(3)***, allows the detection of more than 300,000 tags with less effort and cost than LongSAGE. The sample size increased the ability to reliably detect low-abundance transcripts and replicates, or multiple samples in a single run.

2. Materials

2.1. Capture of Poly-Adenylated mRNA on Paramagnetic Beads

1. 10–100 μg of high-quality total RNA from any eukaryotic sample.
2. Dynabeads Oligo$(dT)_{25}$ (Dynal Biotech Asa, Oslo, Norway).
3. Lysis Buffer: 100 m*M* Tris-HCl, pH 7.5, 500 m*M* LiCl, 10 m*M* ethylenediamine tetraacetic acid (EDTA), 1 % lithium dodecyl sulfate, 5 m*M* dithiothreitol (DTT) (Invitrogen, Carlsbad, CA). Lysis Buffer can be prepared in advance and stored at room temperature, if DTT is excluded and added just prior to use.
4. Wash Buffer A: 10 m*M* Tris-HCl, pH 7.5, 0.15 *M* LiCl, 1 m*M* EDTA, 0.1 % lithium dodecyl sulfate, 10 μg/mL glycogen (Fermentas, Burlington, Canada). Store at −20 °C.
5. Wash Buffer B: 10 m*M* Tris-HCl, pH 7.5, 150 m*M* LiCl, 1 *M* NaCl, 1 % sodium dodecyl sulfate (SDS), 10 μg/mL glycogen. Store at −20 °C.
6. 5X First Strand Buffer: 250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen). Store at −20 °C.

2.2. On-Bead Complementary DNA Synthesis

1. Diethylpyrocarbonate (DEPC) water. Store at room temperature.
2. dNTP mix, 25 m*M* each (Fermentas). Store at −20 °C.
3. 5X First Strand Buffer: 250 m*M* Tris HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen).
4. 0.1 *M* DTT (Invitrogen). Store at −20 °C.
5. SuperScript™ II Reverse Transcriptase (200 U/μL) (Invitrogen). Store at −20 °C.

6. 5X Second Strand Buffer: 100 m*M* Tris-HCl, pH 6.9, 450 m*M* KCl, 23 m*M* $MgCl_2$, 0.075 m*M* β-NAD^+, 50 m*M* $(NH_4)_2SO_4$ (Invitrogen). Store at −20 °C.
7. RNase inhibitor (40 U/μL) (New England Biolabs, Ipswich, MA). Store at −20 °C.
8. *Escherichia coli* DNA ligase (10 U/μL) (Invitrogen). Store at −20 °C.
9. *E. coli* DNA polymerase (10 U/μL) (Invitrogen). Store at −20 °C.
10. *E. coli* RNase H (5 U/μL) (Fermentas). Store at −20 °C.
11. 0.5 *M* EDTA (Bie & Berntsen A-S, Rødovre, Denmark). Store at room temperature.
12. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen. Store at room temperature.
13. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL bovine serum albumin (BSA) (New England Biolabs). Store at −20 °C.
14. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs). Store at −20 °C.

2.3. Digesting Complementary DNA With the Anchoring Enzyme (NlaIII)

1. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5. Store at room temperature.
2. 100X BSA (New England Biolabs). Store at −20 °C.
3. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs). Store at −20 °C.
4. *Nla*III (10 U/μL) (New England Biolabs). Store at −80 °C.
5. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen. Store at −0 °C.
6. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL BSA. Store at −20 °C.

2.4. Ligating Linkers to Bound Complementary DNA

1. DNA synthesis- and protein sequencing-grade water (hereafter referred to as DNA-grade water) (AppliChem, Darmstadt, Germany).
2. Linker1 A:
   5′-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA TCC GAC ATG-3′
   Linker1 B:
   5′-PO_4 TCG GAT ATT AAG CCT AGT TGT ACT GCA CCA GCA AAT CC (Amino C7)-3′
   Linker2 A:
   5′-TTT CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC GTC CGA CAT G-3′
   Linker2 B:
   5′-PO_4 TCG GAC GTA CAT CGT TAG AAG CTT GAA TTC GAG CAG (Amino C7)-3′

(TAG Copenhagen, Copenhagen, Denmark). Dissolve linker oligonucleotides in DNA-grade water to a final concentration of 100 μ*M* and store at −20 °C.
3. 10X T4 DNA Ligase Buffer: 400 m*M* Tris-HCl, 100 m*M* $MgCl_2$, 100 m*M* DTT, 5 m*M* ATP (Fermentas). Store at −20 °C.
4. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH7.5. Store at room temperature.
5. T4 DNA Ligase (5 U/μL) (Fermentas, Burlington, Canada). Store at −20 °C.
6. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL BSA. Store at −20 °C.

2.5. Releasing SAGE Tags Using MmeI

1. 32 m*M* *S*-adenosylmethionine (SAM) (New England Biolabs). Store at −80 °C.
2. DEPC water
3. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs). Store at −20 °C.
4. DNA-grade water (AppliChem, Darmstadt, Germany).
5. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA.
6. *Mme*I (2 U/mL) (New England Biolabs). Store at −20 °C.
7. Phenol:chloroform:isoamyl alcohol (PCI) (25:24:1) saturated with Tris to pH 8.0. (Sigma-Aldrich, St. Louis, MO). Store at 5 °C.
8. 7.5 *M* ammonium acetate (Sigma-Aldrich). Store at room temperature.
9. Glycogen (20 mg/mL). (Fermentas). Store at −20 °C.
10. 100 % ethanol (De Danske Spritfabrikker, Aalborg, Denmark). Store at 5 °C.
11. 70 % ethanol. Store at 5 °C.

2.6. Ditag Formation

1. 3 m*M* Tris-HCl, pH 7.5.
2. 10X T4 DNA Ligase Buffer (Fermentas). Store at −20 °C.
3. DEPC water.
4. T4 DNA Ligase (5 U/μL) (Fermentas). Store at −20 °C.
5. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH7.5. Store at room temperature.

2.7. Attaching Nucleotide Identification Keys During Amplification of Ditags

1. Cold DNA-grade water.
2. 10X PCR Buffer: 100 m*M* Tris-HCl, pH 8.3, 500 m*M* KCl, 15 m*M* $MgCl_2$, 1 % Triton X-100 (Bie & Berntsen A-S, Rødovre, Denmark). Store at −20 °C.
3. Identification and amplification primer mix: 100 μ*M* of each of the two primers in DNA-grade water: 5′-GCCTTGCCAGCCCGCTCAGCAAGCTTCTAACGA TGTACGT-3′ and 5′-GCCTCCCTCGCGCCATCAG**AAG**TGGTGCAGTACA ACTAGGCT. The boldfaced three nucleotides are varied from sample to sample. Store at −20 °C.
4. 25 m*M* $MgCl_2$ (Fermentas). Store at −20 °C.

5. dNTP mix, 25 m*M* each (Fermentas). Store at −20 °C.
6. Taq polymerase (5 U/μL) (Bie & Berntsen A-S, Rødovre, Denmark). Store at −20 °C.
7. 30 % acrylamide/bis (19:1) (AppliChem). Store at 5 °C.
8. Sterile, autoclaved MilliQ water.
9. 50X TAE Buffer: 2 *M* Tris, 1 *M* acetic acid, 0.05 *M* EDTA, pH 8.3. Store at room temperature.
10. *N,N,N′,N′*-tetramethylethylendiamine (TEMED) (Bie & Berntsen A-S, Rødovre, Denmark). Store at 5 °C.
11. Ammonium persulfate: prepare 10 % solution in water and store at −20 °C.
12. Molecular weight markers for gel electrophoresis: GeneRuler™ 100-bp DNA ladder (Fermentas) and 25-bp DNA ladder (Invitrogen), diluted to a final concentration of approx 0.1 μg/μL.
13. Running Buffer (1X TAE Buffer): 40 m*M* Tris, 20 m*M* acetic acid, 1 m*M* EDTA pH 8.3
14. 6X TAE Loading Buffer: 240 m*M* Tris, 120 m*M* acetic acid, 6 m*M* EDTA pH 8.3, 17 % glycerol, bromophenol blue.
15. Staining solution: 25 mL 1X TAE containing 5 μL ethidium bromide 10 mg/mL (Sigma-Aldrich).
16. PCI (25:24:1). Store at 5 °C.
17. 7.5 *M* ammonium acetate. Store at room temperature.
18. Glycogen 20 mg/mL. Store at −20 °C.
19. 100 % ethanol.
20. 70 % ethanol.
21. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5
22. TE: 10 m*M* Tris-HCl, pH 7.5, 1 m*M* EDTA, pH 7.5
23. 50-mL centrifuge tubes.
24. 10-mL glass pipet for PCI extraction (*see* **Note 1**).

2.8. Purifying Ditags

1. 30 % acrylamide/*bis* (19:1).
2. Sterile, autoclaved MilliQ water.
3. 50X TAE Buffer: 2 *M* Tris, 1 *M* acetic acid, 0.05 *M* EDTA, pH 8.3.
4. TEMED.
5. 10 % ammonium persulfate. Store at −20 °C.
6. Molecular weight markers for gel electrophoresis: GeneRuler 100-bp DNA ladder and 25-bp DNA ladder, diluted to a final concentration of 0.1 μg/μL.
7. Running Buffer (1X TAE Buffer): 40 m*M* Tris, 20 m*M* acetic acid, 1 m*M* EDTA pH 8.3.
8. 6X TAE Loading Buffer: 240 m*M* Tris, 120 m*M* acetic acid, 6 m*M* EDTA pH 8.3, 17 % glycerol, bromophenol blue.
9. Staining solution: 25 mL 1X TAE containing 5 μL ethidium bromide 10 mg/mL.

10. PCI (25:24:1).
11. 7.5 *M* ammonium acetate.
12. Glycogen 20 mg/mL.
13. 100 % ethanol.
14. 70 % ethanol.
15. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
16. Spin-X® tubes (Corning Costar Inc., NY).

3. Methods

The method described here produces ample ditag DNA molecules to serve as template DNA for pyrophosphatase-based sequencing by 454-Life Science Corp.

3.1. Capture of Poly-Adenylated mRNA on Paramagnetic Beads

1. Thoroughly resuspend the Oligo$(dT)_{25}$ beads and transfer 100 µL to an RNase-free 1.5-mL tube. Place the tube on a magnetic stand for 1–2 min, and discard supernatant using a pipet.
2. Wash Oligo$(dT)_{25}$ beads by resuspending them in 500 µL of Lysis Buffer. Place the tube in a magnetic stand, but do not remove the supernatant yet.
3. Combine10–50 µg of total RNA (*see* **Note 2**) with Lysis Buffer to a final volume of 1 mL. Carefully remove the supernatant from Oligo$(dT)_{25}$ beads, and immediately add the RNA sample. Incubate the beads and RNA sample by on a rocking platform for 30 min at room temperature.
4. Carefully remove and discard the supernatant from the washed beads, and wash twice with 1 mL Wash Buffer A by placing the tube on magnetic stand for 1–2 min and removing the supernatant between washes. Wash with 1 mL Wash Buffer B and four times with 100 µL 1X First Strand Buffer. After the fourth wash, do not remove supernatant (*see* **Note 3**).

3.2. On-Bead Complementary DNA Synthesis

1. Prepare the first strand complementary DNA (cDNA) reaction mix by combining the following reagents on ice: 18 µL 5X First Strand Buffer, 1 µL RNase inhibitor, 57.2 µL DEPC water, 9 µL 0.1 *M* DTT, and 1.8 µL dNTP mix.
2. Remove the supernatant from the Oligo$(dT)_{25}$ beads and resuspend the beads containing mRNA in the first strand cDNA reaction mix. Mix gently by flicking the tube with a finger. Store the tube at 37 °C for 2 min to equilibrate the reagents. Add 3 µL reverse transcriptase. Mix gently and incubate at 37–42 °C for 1 h. Keep the beads suspended by flicking the tube every 5 min.
3. Place the reaction on ice for 2 min, and add the second strand synthesis reagents in the following order to the tube: 474 µL DEPC water, 150 µL 5X Second Strand Buffer, 6 µL dNTP mix, 5 µL *E. coli* DNA ligase, 20 µL *E. coli* DNA polymerase,

2 μL *E. coli* RNase H. Mix contents by vortexing, and centrifuge the tube briefly (a few seconds at low speed) in a benchtop centrifuge. Incubate the reaction mixture at 16 °C for 2 h, flicking the tube every 5 min to keep the beads suspended.
4. Preheat Wash Buffer C to 75 °C.
5. Place the reaction tube on ice and add 45 μL 0.5 *M* EDTA to stop the reaction. Place the tube on a magnetic stand for 1–2 min and carefully remove supernatant (*see* **Note 4**). Add 750 μL warm Wash Buffer C to inactivate the *E. coli* DNA polymerase. Mix well and heat the sample to 75 °C for 10 min with intermittent mixing to completely inactivate the polymerase. Place the tube on magnetic stand for 1–2 min. and remove supernatant. Wash again with 750 μL Wash Buffer C. Perform the wash quickly to prevent precipitation of SDS, which may trap the beads.
6. Wash sample three times with 750 μL Wash Buffer D and suspend the beads in 750 μL Wash Buffer D. Place the tube on magnetic stand for 1–2 min and carefully remove supernatant.
7. Add 200 μL 1X Buffer 4 to the tube and gently resuspend the beads. Transfer the contents of the tube to a new tube to avoid any traces of exonuclease activity from *E. coli* DNA polymerase. Wash the old tube with 200 μL 1X Buffer 4 and transfer the wash to the tube containing the reaction mix. Place tube on magnetic stand for 1–2 min and remove supernatant.
8. Wash the beads once with 200 μL 1X Buffer 4.

3.3. Cleavage of cDNA With the Anchoring Enzyme (NlaIII)

1. Remove supernatant and resuspend beads in 172 μL LoTE, 2 μL 100X BSA, 20 μL 10X Buffer 4, 6 μL *Nla*III. Incubate for or 1 h at 37 °C. Mix occasionally by flicking the tube. Equilibrate Wash Buffer C to 37 °C to prevent SDS precipitation.
2. After the reaction is complete, place the tube containing the beads on a magnetic stand for 1–2 min and carefully remove and discard the supernatant.
3. Inactivate *Nla*III by washing the tube twice with 750 μL warm Wash Buffer C and wash three times with 750 μL Wash Buffer D.
4. Proceed immediately to ligating linkers to bound cDNA.

3.4. Ligating Linkers to Bound cDNA

Prior to use, the linker oligonucleotides are hybridized and purified by gel electrophoresis to obtain Adaptor A and Adaptor B. The Adaptors are stored at −20 °C in aliquots for single use.

1. Mix the following in two separate tubes: 17 μL Linker1 A, 17 μL Linker1 B, and 4 μL 10X polynucleotide kinase buffer (Adaptor A); 17 μL Linker2 A, 17 μL Linker2 B, and 4 μL 10X polynucleotide kinase buffer (Adaptor B). Place the two tubes in 100 mL boiling water, and incubate at room temperature for 1 h.
2. Load the two LS Adaptor ligation mixes onto a 10 × 8 cm × 0.75 mm 15 % polyacrylamide gel (comb with 10 lanes) alongside a 25-bp DNA ladder. Conduct gel

electrophoresis in 1X TAE running buffer at 120 V for 1 h. Stain gel in 25 mL 1X TAE running buffer containing 5 μL ethidium bromide for 5 min at room temperature and visualize DNA by exposure to ultraviolet (UV) light.

3. Isolate 40-bp bands for LS Adaptors A and B using a clean scalpel. Transfer gel pieces to two 0.5-mL Eppendorf tubes that have been punctured at the bottom by a 16-G needle. Insert the small tubes into 1.5-mL Eppendorf tubes and centrifuge at maximum speed in a benchtop centrifuge for 30–60 s. The gel pieces are forced through the small hole and crushed to bits in the process.
4. Add 375 μL LoTE and 125 μL 7.5 *M* ammonium acetate to the crushed gel in each tube. Mix by vortexing and elute adaptors by incubation over night at 4 °C. Transfer all of the contents to 2-mL Spin-X filter tubes and centrifuge at maximum speed in a benchtop centrifuge for 30 s. Add 2 μL glycogen and 1500 μL 100 % ethanol to each tube and vortex briefly. Store tube at −80 °C (or better, on dry ice) for 30 min, centrifuge at maximum speed in a benchtop centrifuge for 30 min, and remove and discard supernatant. Wash pellets in 1 mL 70 % ethanol. Leave tube on the bench top with the lid open and air-dry pellets for a minimum of 10 min. Resuspend each pellet in 100 μL TE and determine LS Adaptor concentration by absorption at 260 nm.
5. To ligate adaptors to immobilized cDNA, place tube with beads on magnetic stand for 1–2 min and carefully remove supernatant. Wash the cDNA containing beads twice with 150 μL of 1X T4 DNA Ligase Buffer. After the final wash, resuspend beads in 100 μL 1X T4 DNA Ligase Buffer and divide the sample into two new tubes labeled A and B. Be careful to divide the Oligo$(dT)_{25}$ beads while they are homogenously suspended.
6. Wash each tube once with 50 μL 1X T4 DNA Ligase Buffer. Resuspend the beads in 50 μL 1X T4 DNA Ligase Buffer and place the tubes on a magnetic stand for 1–2 min and carefully remove the supernatant. Transfer the tubes to ice and immediately add 1 μL LS Adaptor A or B (60 ng/μL), 14.5 μL LoTE, 2 μL 10X T4 DNA Ligase Buffer. Resuspend the beads by flicking each tube and heat at least 2 min at 50 °C. Cool the tube to room temperature for 15 min and place on ice. Add 2 μL T4 DNA Ligase to each tube and mix well. Incubate overnight at 16 °C.
7. The following day, wash each tube three times with 500 μL of Wash Buffer D.

3.5. Releasing SAGE Tags Using MmeI

1. Prepare 10X SAM by adding 1 μL 32 m*M* SAM to 79 μL DEPC water.
2. Prepare 1X Buffer 4/1X SAM by combining 80 μL 10X Buffer 4, 720 μL DNA-grade water, and 1 μL 32 m*M* SAM.
3. Place the two tubes on a magnetic stand for 2 min and remove supernatant. Wash each tube twice with 200 μL 1X Buffer 4/1X SAM. Carefully remove and discard the supernatant and place tubes on ice. Add the following to each tube: 70 μL LoTE, 10 μL 10X Buffer 4, 10 μL 10X SAM, 10 μL *Mme*I. Incubate tubes at 37 °C for 2.5 h with occasional gentle mixing.

4. Place tubes on magnetic stand for 2 min. This time, do not discard the supernatant (*see* **Note 5**). Carefully remove the supernatant from each tube and pool supernatants in a new tube. Add 100 µL of LoTE to yield a total volume of 300 µL.
5. Add 300 µL PCI (25:24:1) to tube and vortex thoroughly. Centrifuge for 5 min at room temperature and at maximum speed. Transfer 300 µL of the upper aqueous phase to a new tube. Transfer 200 µL from this tube to a new tube. The remaining 100 µL are used as a negative control for the ligase reaction (no ligase). Add 100 µL DEPC water to the no-ligase reaction to yield a final volume of 200 µL.
6. To each tube (200 µL sample and 200 µL negative control), add 133 µL 7.5 *M* ammonium acetate, 3 µL glycogen, and 1 mL of 100 % ethanol. Mix vigorously.
7. Store tubes at −80 °C for a minimum of 30 min and centrifuge at maximum speed in a benchtop centrifuge for 30–40 min at 4 °C. Carefully remove supernatant without disturbing the pellet from each tube and discard it.
8. Wash each pellet twice with 1 mL of cold 70 % ethanol. After the final wash, centrifuge each tube again to collect any residual ethanol to the bottom of the tube. Remove the ethanol using a pipet and leave the tube on the bench top for 5–10 min with the lid open to remove any traces of ethanol.
9. Resuspend the sample in 4 µl LoTE and the no-ligase control pellet in 2 µL LoTE and incubate at 37 °C for 10–15 min to ensure solubilization.

3.6. Ditag Formation

1. Prepare 2X ditag reaction in a sterile microcentrifuge tube on ice by combining 1.5 µL 3 m*M* Tris-HCl, pH 7.5, 0.9 µL 10X T4 DNA Ligase Buffer, 1.1 µL DEPC water, and 1 µL T4 DNA Ligase.
2. Prepare 2X Negative Control in a sterile microcentrifuge tube on ice by combining 2.25 µL 3 m*M* Tris-HCl, pH 7.5, 0.75 µL 10X T4 DNA Ligase Buffer, and 0.75 µL DEPC water.
3. Add 4 µL of 2X Ditag Reaction Mix to the tag solution from above and add 2 µL of 2X Negative Control Mix to the no-ligase control. Incubate tubes overnight at 16 °C.

3.7. Attaching Nucleotide Identification Keys During Amplification of Ditags

To optimize PCR conditions, a test PCR is performed using dilutions of ditags (1:40, 1:80, 1:160, 1:320, 1:640). Ligated LS Adaptors are used as positive control (1:40). For negative controls, both a no-template and the no-ligase control are amplified.

1. For each PCR reaction mix the following on ice: 1 µL template (diluted ditags and controls), 36.5 µL cold DNA-grade water, 5 µL 10X PCR Buffer, 1 µL LS DTP primer mix, 5 µL $MgCl_2$, 1 µL dNTP mix, and 0.5 µl Taq polymerase (*see* **Note 6**).

2. Perform PCR according to the following procedure: initial denaturation at 94 °C for 1 min, followed by 28 cycles of denaturation at 94 °C for 30 s, annealing at 53.5 °C for 1 min, and elongation at 70 °C for 1 min (*see* **Note 7**).
3. Load 5 μL of each PCR reaction onto a 10 × 8 cm × 0.75 mm 15 % polyacrylamide gel (comb with 10 lanes). Conduct gel electrophoresis in 1X TAE running buffer for 45 min at 100 V and 45 min at 120 V. Stain the gel for 5 min in 25 mL 1X TAE containing 5 μL ethidium bromide (*see* **Fig. 1**).
4. Select the most appropriate ditag dilution and perform 5–10 50-μL PCR reactions at these conditions. Pool the resulting PCR reactions and place on ice. Analyze 5 μL on a 10 × 8 cm × 0.75 mm 15 % polyacrylamide gel alongside a 25-bp DNA ladder. Conduct gel electrophoresis in 1X TAE running buffer for 45 min at 100 V and 45 min at 120 V. Stain the gel for 5 min in 25 mL 1X TAE containing 5 μL ethidium bromide. Visualize bands by exposure to UV light (*see* **Fig. 2**).
5. Store amplified ditags at −40 °C or proceed to Isolation of ditags.

3.8. Purifying Ditags

1. Extract amplified ditags by adding an equal amount of PCI. Vortex well, centrifuge for 10 min at room temperature, and transfer 400 μL upper aqueous phase to a 2-mL tube (*see* **Note 8**).
2. Add 100 μL 7.5 *M* ammonium acetate, 5 μL glycogen, and 1250 μL cold 100 % ethanol. Mix vigorously. Store tubes at −40 °C for a minimum of 30 min or overnight. Centrifuge at 12,000*g* for 30 min. Carefully remove and discard supernatant. Wash pellet twice with 1 mL cold 70 % ethanol. Remove ethanol and air-dry pellet for 15–20 min.

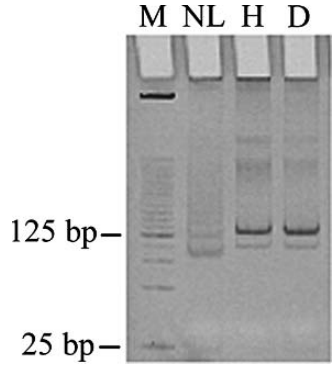


Fig. 1. Nucleotide identification key-tagged, amplified serial analysis of gene expression ditags. M, 25-bp DNA marker; NL, no-ligase control reaction. H and D are ditags derived from mRNA from potato tubers at harvest (H) and 60 d postharvest (D). Five out of 300 μL PCR product was loaded on the gel.

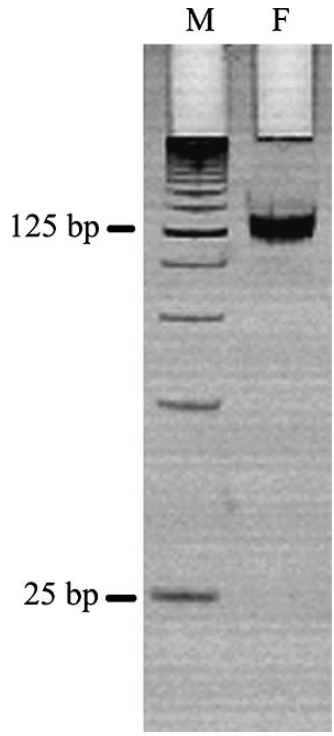


Fig. 2. Purified ditags suitable for sequencing. M, 25-bp DNA marker and F, final pooled ditags from potato tubers after TAE-polyacrylamide gel electrophoresis purification. Two microliters out of a total volume of 20 μL was loaded on the gel.

3. Resuspend pellets in 100 μL LoTE, and incubate tubes at 37 °C for 5–10 min to aid in solubilization. Centrifuge for 5 min at maximum speed. Transfer supernatant to a new tube.
4. Load the sample in a single lane of a 1.5-mm 12 % TAE-polyacrylamide gel electrophoresis (PAGE) gel. Conduct gel electrophoresis in 1X TAE running buffer at 100 V for 90 min. Stain the gel for 5 min in 25 mL1X TAE containing 5 μL ethidium bromide. Visualize bands by exposure to UV light. Excise 130-bp product using a clean scalpel.
5. To crush the excised gel piece, place it in a 0.5-mL tube that has been punctured at the bottom with a 16-G needle. Place that tube in a 1.5-mL tube and centrifuge for 2 min at maximum speed. Elute the DNA by adding 375 μL LoTE and 125 μL 7.5 *M* ammonium acetate to the crushed gel in the large tube and incubate overnight at 4 °C.
6. Transfer the entire content to 2-mL Spin-X filter tubes and centrifuge at maximum speed in a benchtop centrifuge for 30 s. Add 2 μL glycogen and 1500 μL 100 %

ethanol to each tube and vortex briefly. Store tube at −80 °C (or better, on dry ice) for 30 min, centrifuge at maximum speed in a benchtop centrifuge for 30 min, and remove and discard supernatant. Wash pellets in 1 mL 70 % ethanol. Leave tube on the bench top with the lid open and air-dry pellets for a minimum of 10 min. Resuspend each pellet in 20 μL TE and estimate concentration, e.g., by absorption at 260 nm or dot blot (*see* **Note 10**), and check the integrity and purity prior to shipment to 454 Life Science Corporation for sequencing by running of a 12 % TAE-PAGE (*see* **Fig. 2**).

4. Notes

1. Glass pipets are recommended when working with PCI. Many types of plastic are dissolved in aggressive organic solvents.
2. 50–100 μg of total RNA is preferable. It is imperative that undegraded RNA be used.
3. It is important that the oligo$(dT)_{25}$ beads do not dry out.
4. The oligo$(dT)_{25}$ beads might stick to the tube, so be very careful when removing the supernatant. After the addition of the 75 °C warm Wash Buffer C, the beads will no longer stick to the tube.
5. During *Mme*I digestion, the cDNA is released from the oligo$(dT)_{25}$ beads. Therefore, the supernatant contains the tags and must not be discarded.
6. DNA-grade water should be added to each PCR tube followed by the templates. A PCR Master Mix can be prepared by adding, in the following order, 10X PCR Buffer; identification and amplification primer mix (175 ng/μL); $MgCl_2$ (25 m*M*); dNTP mix (25 m*M* of each); and Taq polymerase (5 U/μL). Mix by pipetting up and down twice using a 200-μL pipet. Always keep PCR reaction mix and PCR tubes on ice.
7. Before placing the PCR reactions in the heat block, preheat it to approx 80 °C.
8. The upper phase contains the ditags.
9. For smaller or larger samples, decrease or increase the amount proportionally.
10. A dot blot can be used to estimate the concentration of the ditag sample. Use 1 μL ditag sample and make the following dilutions: 1:5, 1:25, and 1:125. Add 1 μL of stain (1 μL ethidium bromide + 10 mL DNA-grade water) to 4 μL of sample dilutions and to 4 μL of each standard solution (20 ng DNA, 10 ng DNA, 5 ng DNA, and 2.5 ng DNA). Pipet 5-μL droplets onto household film, visualize under UV light, and compare samples with standards.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
2. Gowda, M., Jantasuriyarat, C., Dean, R. A., and Wang, G. L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol.* **134**, 890–897.

3. Margulies, M., Egholm, M., Altman. W. E., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
4. Saha, S. Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512.

7

High-Resolution, Genome-Wide Mapping of Chromatin Modifications by GMAT

Tae-Young Roh and Keji Zhao

Summary

One major postgenomic challenge is to characterize the epigenomes that control genome functions. The epigenomes are mainly defined by the specific association of nonhistone proteins with chromatin and the covalent modifications of chromatin, including DNA methylation and posttranslational histone modifications. The in vivo protein-binding and chromatin-modification patterns can be revealed by the chromatin immunoprecipitation assay (ChIP). By combining the ChIP assays and the serial analysis of gene expression (SAGE) protocols, we have developed an unbiased and high-resolution genome-wide mapping technique (GMAT) to determine the genome-wide protein-targeting and chromatin-modification patterns. GMAT has been successfully applied to mapping the target sites of the histone acetyltransferase, Gcn5p, in yeast and to the discovery of the histone acetylation islands as an epigenetic mark for functional regulatory elements in the human genome.

Key Words: Chromatin; epigenetics; ChIP; histone; epigenome.

1. Introduction

The expression patterns of eukaryotic genomes are controlled by their epigenomes mainly on the level of chromatin modifications. The basic structural unit of chromatin is a nucleosome that is formed by wrapping approx 146 bp of DNA around a core of eight histone molecules. Various posttranslational histone modifications regulate gene activity by modifying directly or indirectly chromatin structure and accessibility ***(1–3)***. The patterns of histone modifications of a specific locus can be determined by chromatin immunoprecipitation

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

(ChIP) assay, which involves stabilization of the histone modifications by formaldehyde cross-linking and immunoprecipitation of a specifically modified histone with its associated DNA using antibodies against the specific modification. The associated DNA is purified and analyzed by PCR to detect the presence of a DNA sequence. Large-scale analysis of the ChIP DNA can be performed on DNA microarrays using the ChIP DNA as probes (ChIP-on-chip). This method depends on the preselected sequences on the arrays. We have developed an unbiased method to determine the genome-wide chromatin modifications of any genomes with known genomic sequences by combining the ChIP assay with the serial analysis of gene expression (SAGE) protocol *(4–6)*, which we have named the genome-wide mapping technique (GMAT) *(7)*. The method is illustrated in **Fig. 1**. Following cross-linking living cells with formaldehyde to stabilize the nucleosome and histone tail modifications, chromatin is fragmented to 300–500 bp by sonication. ChIP assay is performed to purify the modified histones with bound DNA using specific antibodies against modified histone tails. After reverse cross-linking, a biotinylated universal linker is ligated to the DNA ends. Treatment with *Nla*III enzyme cleaves the majority of the ChIP DNA. Linkers with the *Mme*I recognition site are ligated to the *Nla*III-cut DNA ends. Following the LongSAGE protocol, short DNA sequence tags (approx 21 bp) are isolated by *Mme*I digestion and concatemerized to 500- to 1000-bp fragments that are cloned into a sequencing vector for sequence analysis. Based on the sequence information, the 21-bp tags are mapped onto the genome, and the detection frequency of a tag represents the level of the histone modification. GMAT has been successfully applied to mapping the genome-wide histone acetylation patterns in the yeast and human genomes *(7–8)*. The distribution of histone H3 acetylation is mapped on yeast chromosome III and human chromosome 12 as an example **(Fig. 2)**. These analyses indicate that histones in the promoter regions are highly acetylated in both yeast and human **(Fig. 3)**. Histone acetylation is an epigenetic mark for functional chromatin and/or transcription regulatory elements *(8)*.

2. Materials

2.1. Yeast Culture and Chromatin Preparation

1. YPD medium: 1% yeast extract, 2% peptone, 2% dextrose.
2. 37% formaldehyde.
3. 1*M* glycine.
4. 1X phosphate-buffered saline (PBS): NaCl (9 g/L), Na_2HPO_4 (0.775 g/L), KH_2PO_4 (0.165 g/L), pH 7.4.

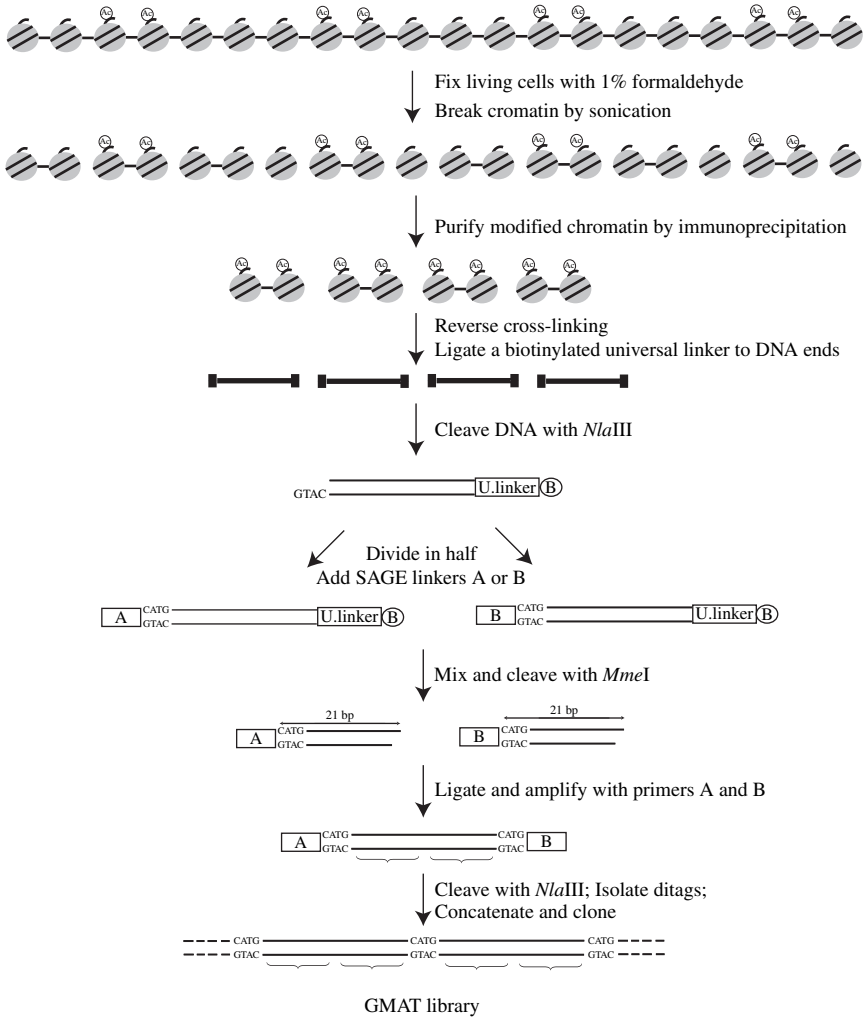


Fig. 1. Schematic illustration of the genome-wide mapping technique (GMAT).

5. Lysis buffer : 50 m*M* HEPES, pH 7.5, 140 m*M* NaCl, 1 m*M* ethylenediamine tetraacetic acid (EDTA), 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% sodium dodecyl sulfate (SDS), 10 m*M* Na-butyrate, 1X proteinase inhibitor cocktail, and 0.1 m*M* fresh phenylmethylsufonyl fluoride (PMSF).
6. Glass beads (425–600 μm, Acid-washed; Sigma-Aldrich, St. Louis, MO)
7. 1X TE: 10 m*M* Tris-HCl, pH 7.4, 1 m*M* EDTA.
8. 10% SDS.

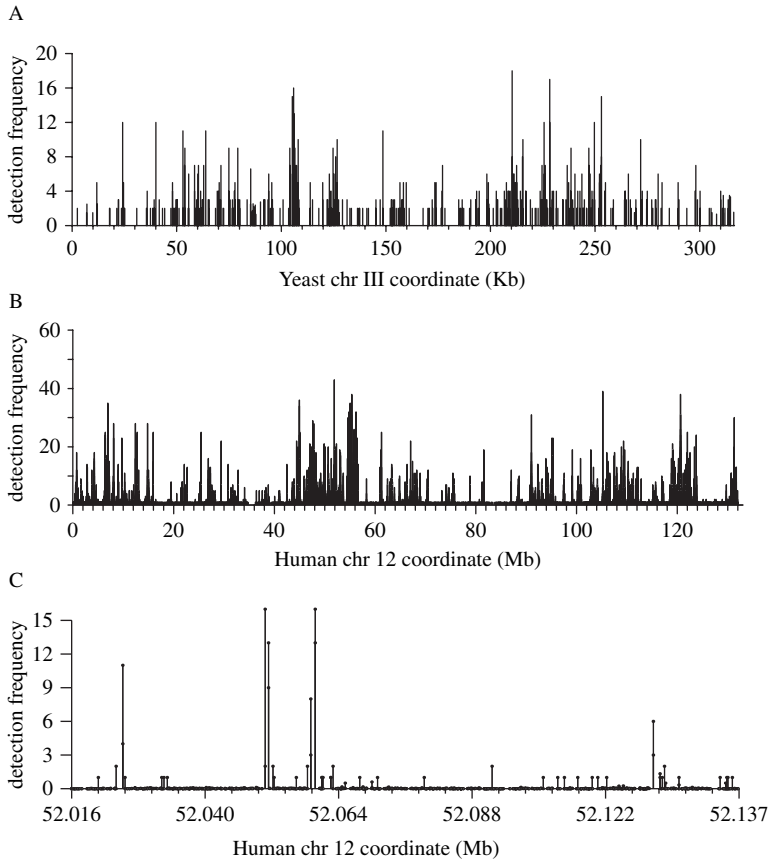


Fig. 2. Histone acetylation map. Distribution of diacetylated K9/K14 histone H3 on yeast chromosome III **(A)** and human chromosome 12 **(B,C)**. The normalized tag count represents the number of times that a particular tag was detected in the genome-wide mapping technique library divided by the number of hits of the tag sequence in the genome.

2.2. Chromatin Immunoprecipitation

1. Radioimmunoprecipitation assay (RIPA) buffer: 1X TE, 0.1% SDS, 0.1% sodium deoxycholate, and 1.0% Triton X-100.
2. nProtein A Sepharose 4 Fast Flow (GE healthcare Bio-Sciences, Uppsala, Sweden Pharmacia).
3. Affinity-purified specific antibody.

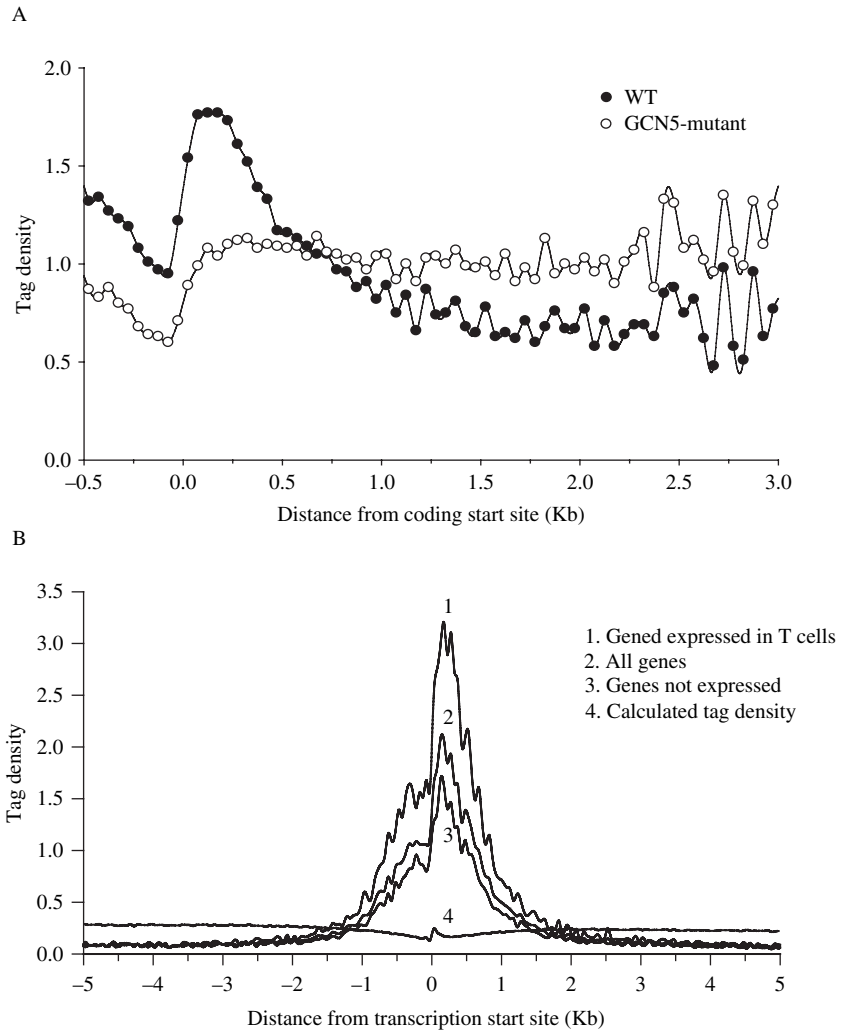


Fig. 3. Distribution of tags in the promoter regions and along the gene body region. Tag distribution was calculated according to the relative distance to ATG start codons of 6040 yeast genes (**A**) and to transcription start sites of 21,355 human genes. Tag density was obtained by normalizing the total number of detected tags to the number of expected *Nla*III sites in a 50-bp window.

4. 5 *M* NaCl stock solution.
5. LiCl buffer: 1X TE, 0.25 *M* LiCl, 0.5% NP-40, 0.5% sodium deoxycholate.
6. 20 mg/mL glycogen.
7. 1 *M* Tris-HCl buffer, pH 8.0 saturated phenol/chloroform (1:1).
8. 3 *M* sodium acetate, pH 5.2.
9. Ethanol.

2.3. Enzymes and Reagents

1. Proteinase K: 10 mg/mL (Invitrogen, Carlsbad, CA).
2. Klenow enzyme: 5 U/μL, (New England Biolabs, Ipswich, MA).
3. T4 DNA ligase: 400 U/μL, (New England Biolabs, Ipswich, MA).
4. PfuTurbo DNA polymerase: 2.5 U/μL, (Stratagene, La Jolla, CA).
5. *Nla*III: 10 U/μL, (New England Biolabs, Ipswich, MA).
6. *Mme*I: 2 U/μL, (New England Biolabs, Ipswich, MA).
7. Dynabeads M-280 Streptavidin and magnetic stand (Invitrogen, Carlsbad, CA).
8. 7.5 *M* ammonium acetate
9. Zeocin (Invitrogen, Carlsbad, CA).
10. pZErO-1 plasmid (Invitrogen, Carlsbad, CA).
11. Electromax DH10Bs (Invitrogen, Carlsbad, CA).

2.4. Linker and Primer Sequences

1. Linker WL1: 5′-[biotin]GCGGTGACCCGGGAGATCTGAATTC-3′ (polyacrylamide gel electrophoresis [PAGE]-purified).
2. Linker WL2: 5′-GAATTCAGATC-3′ (PAGE-purified).
3. Linker 1A: 5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATATCCGACATG-3′(PAGE-purified).
4. Linker 1B: 5′-TCGGATATTAAGCCTAGTTGTACTGCACCAGCAAATCC[aminomod.C7]-3′(PAGE-purified).
5. Linker 2A: 5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGTCCGACATG-3′(PAGE-purified).
6. Linker 2B: 5′-TCGGACGTACATCGTTAGAAGCTTGAATTCGAGC-AG[aminomod.C7]-3′(PAGE-purified).
7. Primer 1: 5′-[dual biotin]GTGCTCGTGGGATTTGCTGGTGCAGTACA-3′ (PAGE-purified).
8. Primer 2: 5′-[dual biotin]GAGCTCGTGCTGCTCGAATTCAAGCTTCT-3′ (PAGE-purified).

2.5. Data Analysis

1. SAGE2000 software ver.4.5 (from www.sage.org).
2. Microsoft Access or other compatible database program.

3. Methods

3.1. Preparation of Yeast Chromatin

1. Grow 200 mL of yeast in YPD medium at 30 °C until the optical density (OD) at 600 nm reaches about 1.0.
2. Add 5.4 mL of 37 % formaldehyde to a final concentration of 1 % and incubate for 15 min at room temperature. To stop the reaction, add 15 mL of 1 *M* glycine and continue the incubation for 5 min at room temperature.
3. Harvest the cells by centrifugation at 1600*g* for 3 min at 4 °C.
4. Wash the pellet twice with 20 mL of 1X PBS.
5. Resuspend the pellets in 4 mL of Lysis buffer.
6. Break the cells with 2 mL of glass beads by vigorous vortexing for 10 min.
7. Remove the cell debris and glass beads by centrifugation at 1600*g* for 3 min at 4 °C.
8. Dispense the lysates equally into two 15-mL plastic centrifuge tubes.
9. Shear chromatin by sonication with 10 30-s pulses at the maximum setting with 30-s intervals in an ice-water bath (*see* **Note 1**).
10. Centrifuge at full speed for 10 min at 4 °C in a microcentrifuge. Store the supernatant in 1-mL aliquots at −80 °C.
11. To determine the average size and concentration of chromatin, add 150 μL of 1X TE, 5 μL of 10 % SDS, and 10 μL of 10 mg/mL proteinase K to 50 μL of the chromatin solution and incubate overnight at 65 °C. Purify the DNA by phenol/chloroform extraction (twice) and ethanol precipitation (*see* **Note 2**).

3.2. Chromatin Immunoprecipitation

1. To a 15-mL centrifuge tube, add 8 mL of RIPA buffer containing 140 m*M* NaCl and 400 μL of protein A sepharose beads (bed volume), and 1 mL (approx0.2 mg) of the chromatin lysate (*see* **Note 3**).
2. Incubate the mixture at 4 °C with rotation for 1 h.
3. Centrifuge at 1000*g* for 3 min at 4 °C.
4. Transfer the chromatin supernatant to a new tube containing 50 μL of fresh protein A sepharose beads (bed volume). Add 5 μg of the affinity-purified anti-acetylated histone H3 antibody. Incubate overnight at 4 °C with rotation (*see* **Note 4**).
5. Centrifuge at 1000*g* for 3 min at 4 °C. Discard the supernatant. Wash beads twice with 5 mL of RIPA buffer containing 140 m*M* NaCl by rotating for 10 min at room temperature.
6. Centrifuge as in **step 5**. Wash beads twice with 5 mL of RIPA buffer containing 300 m*M* NaCl by rotating for 10 min at room temperature.
7. Centrifuge as in **step 5**. Wash beads twice with 5 mL of LiCl buffer by rotating for 10 min at room temperature.

8. Centrifuge as in **step 5**. Wash beads twice with 5 mL of 1X TE by rotating for 10 min at room temperature.
9. Resuspend beads in 200 μL of 1X TE. Add 5 μL of 10 % SDS, and 10 μL of 10 mg/mL proteinase K and incubate overnight at 65 °C. Also treat 100 μL of the input chromatin in the same way.
10. Centrifuge for 10 s in a microcentrifuge. Transfer the supernatant to a new Eppendorf tube. Wash the beads with 100 μL of 1X TE. Centrifuge and combine the supernatants.
11. Extract with phenol/chloroform twice.
12. Precipitate the DNA with 1 μL of 20 mg/mL glycogen (carrier), 30 μL of 3 *M* sodium acetate, pH 5.2 and 700 μL of ethanol.
13. Wash pellet with once 70 % ethanol, air-dry briefly, and resuspend DNA in 20 μL of 1X TE.

3.3. Klenow Treatment and Linker Ligation

1. To 13 μL of DNA, add 3 μL of 1 m*M* dNTPs, 2 μL of 10X Klenow buffer (100 m*M* Tris-HCl, pH 7.5; 50 m*M* $MgCl_2$; 75 m*M* dithiothreitol [DTT]), and 2 μL of Klenow enzyme (10 U). Incubate the mixture at 37 °C for 20 min.
2. Stop the reaction by adding 80 μL of 1X TE. Extract with phenol/chloroform. Precipitate the DNA by adding 10 μL of 3 *M* sodium acetate, pH 5.2, 1 μL of 20 mg/mL glycogen, and 250 μL of ethanol.
3. Wash the pellet with 70 % ethanol.
4. Air-dry briefly and resuspend the DNA in 19 μL of 1X TE.
5. Anneal the 5'-biotinylated WL1 and nonbiotinylated WL2: mix 20 μL of 40 μ*M* WL1, 20 μL of 40 μ*M* WL2, 55 μL of H_2O, and 5 μL of 10X NEB buffer 2. Heat at 95 °C for 5 min, then cool down to 4 °C slowly (about 2 h). Store the mixture at −20 °C (final concentration: 8 μ*M*).
6. Add 5 μL of 10X T4 DNA ligase buffer (500 m*M* Tris-HCl, pH 7.5; 100 m*M* $MgCl_2$; 100 m*M* DTT; 250 μg/mL bovine serum albumin [BSA]), 5 μL of the annealed WL1(biotin)+WL2, 19 μL of H_2O, and 2 μL of T4 DNA ligase (400 U/μL) to the 19 μL of DNA from **step 4**.
7. Incubate at 14 °C for 16 h.
8. Purify the DNA by phenol/chloroform extraction and ethanol precipitation.
9. Resuspend the DNA in 10 μL of 1X TE.
10. Amplify the ligated DNA: mix 8 μL of DNA, 5 μL of 10X Pfu buffer, 5 μL of 1 m*M* dNTPs, 5 μL of 10 μ*M* biotinylated WL1, 26 μL of H_2O, and 1 μL of 2.5 U/μL Pfu enzyme. Incubate the reaction mixture at 74 °C for 10 min (*see* **Note 5**). Then, perform 15 cycles of the following: 94 °C, 30 s; 58 °C, 30 s; 74 °C, 45 s (*see* **Note 6**).
11. Purify the amplified DNA (about 3 μg) by phenol/chloroform extraction and ethanol precipitation. Resuspend the DNA in 30 μL of 1X TE.

3.4. NlaIII Digestion and MmeI Linker Ligation

1. Mix the following: 25.6 μL of DNA, 3 μL of 10X *Nla*III buffer, 0.4 μL of 100X BSA, 1 μL of *Nla*III (10 U).
2. Incubate at 37 °C for 2 h.
3. Add 70 μL of 1X TE, extract with equal volume of phenol/chloroform, and precipitate with ethanol.
4. Wash with 70 % ethanol, air-dry briefly, and resuspend the DNA in 8 μL of 1X TE and 50 μL of H_2O.
5. Add 8 μL of 10X T4 DNA ligase buffer to the 58 μL of DNA and dispense the DNA into two tubes (tube 1 and tube 2, 33 μL each).
6. Anneal Linker 1A and Linker 2A to their complementary Linker 1B and Linker 2B, respectively, as following. Mix 20 μL of 40 μ*M* Linker 1A (or Linker 2A), 20 μL of 40 μ*M* Linker 1B (or Linker 2B), 55 μL of H_2O, and 5 μL of 10X NEB buffer 2. Heat at 95 °C for 5 min, then cool down to 4 °C slowly (about 2 h). Store the mixture at −20 °C (final concentration: 8 μ*M*) (*see* **Note 78**).
7. Add 5 μL of 8 μ*M* annealed Linker 1 to tube 1 and 5 μL of 8 μ*M* annealed Linker 2 to tube 2 and heat the tubes at 50 °C for 2 min.
8. Incubate at room temperature for 10 min.
9. Add 2 μL of T4 DNA ligase to each tube.
10. Incubate at 16 °C overnight.
11. Combine the ligation mixture and purify the DNA by phenol/chloroform extraction and ethanol precipitation.

3.5. Binding to Streptavidin Beads and MmeI Digestion

1. Resuspend the DNA in 20 μL of 1X TE.
2. Add 100 μL of Dynabeads M-280 Streptavidin slurry to a new Eppendorf tube and use a magnetic stand to remove supernatant.
3. Wash the beads with 200 μL of binding buffer (1X TE, 1 *M* NaCl).
4. Add 100 μL of the binding buffer and 20 μL of the DNA to the beads.
5. Incubate for 15 min at room temperature with occasional mixing.
6. Discard the supernatant and wash beads twice with 200 μL of the binding buffer.
7. Wash beads once with 200 μL of 1X TE.
8. Immediately digest bead-bound DNA by adding 86 μL of 1X TE, 158 μL of H_2O, 30 μL of 10X *Mme*I buffer, 3 μL of 5 m*M S*-adenosylmethionine (SAM), 3 μL of 100X BSA, and 20 μL of *Mme*I (2 U/μL).
9. Incubate at 37 °C for 3 h with occasional mixing.
10. Collect the supernatant and purify the DNA by phenol/chloroform extraction.
11. Precipitate the DNA with 2 μL of 20 mg/mL glycogen, 30 μL of 3 *M* sodium acetate, pH 5.2, and 825 μL of ethanol. Incubate in dry ice for 10 min. Centrifuge for 15 min at 4 °C in a microcentrifuge.
12. Wash the pellet once with 70 % ethanol.

13. Resuspend the DNA in 9 μL of 1X ligase buffer and add 1 μL of T4 DNA ligase.
14. Incubate at 16 °C overnight.

3.6. PCR Amplification of Ditags

1. Add 90 μL of 1X TE to the 10 μL of ligation mixture.
2. Prepare the following PCR mixture: 100 μL of 10X Pfu buffer, 100 μL of 1 m*M* dNTPs, 50 μL of dimethylsulfoxide (DMSO), 35 μL of 10 μ*M* primer 1, 35 μL of 10 μ*M* primer 2, 640 μL of H_2O, 20 μL of ligation mixture, 20 μL of 2.5 U/μL pfu enzyme. Aliquot 50 μL of reaction mixture to each tube.
3. Incubate the reaction mixture at 74 °C for 10 min (*see* **Note 5**)
4. Run 25 cycles as follows: 94 °C, 20 s; 55 °C, 30 s; 74 °C, 40 s. Finally, incubate for 5 min at 74 °C.
5. Pool reactions into three tubes (330 μL each) and purify the DNA by phenol/chloroform extraction.
6. Precipitate the DNA by adding 33 μL of 3 *M* sodium acetate and 800 μL of ethanol. Incubate for 5 min in dry ice. Centrifuge for 15 min at 4 °C.
7. Wash twice with 70 % ethanol.
8. Resuspend the DNA in 350 μL of 1X TE.

3.7. NlaIII Digestion and Isolation of Ditags

1. Mix 320 μL of DNA, 40 μL of 10X *Nla*III buffer, 4 μL of 100X BSA, and 40 μL of *Nla*III (10 U/μL)
2. Incubate for 2 h at 37 °C. Stop the reaction by adding 400 μL of 1X TE containing 2 *M* NaCl.
3. Prepare four tubes labeled with A through E. Add 0.1 mL (1 mg) of Streptavidin beads (Dynabeads M-280) to Tube A, and 20 μL (0.2 mg) of Dynabeads M280 to B, C, D, and E, respectively.
4. Remove supernatant by magnet and resuspend the beads in 200 μL of wash buffer (1X TE containing 1 *M* NaCl and 1X BSA).
5. Remove the wash buffer from A and then add the 800 μL of reaction mix to the beads.
6. Mix at room temperature for 15 min.
7. Remove the wash buffer from B and then transfer the supernatant from A to B and mix for 10 min.
8. Remove the wash buffer from C and then transfer the supernatant from B to C and mix for 10 min.
9. Remove the wash buffer from D and then transfer the supernatant from C to D and mix for 10 min.
10. Remove the wash buffer from E and then transfer the supernatant from D to E and mix for 10 min.
11. Collect the supernatant from E and dispense to four tubes (200 μL each).
12. Extract with equal volume of phenol/chloroform.

13. To the 200 μL of DNA, add 66 μL of 7.5 *M* ammonium acetate, 2 μL of 20 mg/mL glycogen, and 825 μL of ethanol.
14. Place in dry ice for 10 min and spin for 15 min at 4 °C.
15. Wash pellets with ice-cold 75 % ethanol.
16. Resuspend the pellets in 30 μL of 1X TE and add 10 μL of 5X BPB loading buffer (1X TAE).
17. Load 6 μL per lane onto a 12 % polyacrylamide gel (7 cm × 8 cm × 1 mm, 15 wells).
18. Run the gel at 100 V for 1 h until the BPB is at the bottom. Use ϕX174/HinfI DNA as a size marker.
19. Stain the gel in 50 mL of H_2O containing 0.5 μg/mL ethidium bromide for 2 min.
20. Cut out the band at around 38 bp, which is close to the 40-/42-bp doublets in the marker DNA.
21. Put the gel slices into two 0.5-mL tubes with a 21-G needle hole at the bottom.
22. Put the 0.5-mL tube in a 2-mL microcentrifuge tube and spin at full speed for 2 min. The gel slices are smashed into the 2-mL tubes through the holes.
23. Remove the 0.5-mL tube and add 250 μL of 1X TE and 50 μL of 7.5 *M* ammonium acetate into each 2-mL tube.
24. Shake at 37 °C for 2 h (or overnight).
25. Transfer the gels and sups into a SNAP spin column (Invitrogen) and spin the eluate into a microcentrifuge tube (2 min at full speed).
26. Dispense the eluate into three Eppendorf tubes (200 μL each) and precipitate by adding to each tube 2 μL of 20 mg/mL glycogen, 66 μL of 7.5 *M* ammonium acetate, and 840 μL of ethanol.
27. Incubate in dry ice for 10 min and spin for 15 min at 4 °C. Wash the pellets twice with ice-cold 75 % EtOH. Briefly air-dry and resuspend the pellets in 8 μL of 1X TE.

3.8. Concatenation of Ditags

1. Add 1 μL of 10X T4 DNA ligase buffer and 1 μL of T4 DNA ligase (400 U/μL) to the 8 μL of DNA.
2. Incubate the mixture at 16 °C for 90 min.
3. Check the ligation efficiency by loading 1 μL of the ligation mix onto 1.4 % agarose/1X TAE gel. Run at 200 V for 10 min (*see* **Note 9**).
4. Stop the reaction by adding 2.5 μL of 5X BPB loading buffer containing 0.1 *M* EDTA.
5. Heat at 65 °C for 5 min. Then, load all of the ligation mixture onto the 1.4 % gel. Run at 120 V for about 1 h.
6. Cut out the gel slices ranging from 500 to 2000 bp.
7. Isolate the DNA using Qiagen gel-extraction kit.
8. Elute the DNA with 200 μL of 1X TE from the QIAEX II beads. Extract with phenol/chloroform.

9. Precipitate the DNA with 2 μL of 20 mg/mL glycogen, 20 μL of 3 *M* sodium acetate, pH 5.2, and 500 μL of ethanol. Incubate in dry ice for 5 min. Spin for 15 min at 4 °C.
10. Wash the pellet twice with 70 % ethanol.
11. Resuspend the pellet in 14 μL of 1X TE.

3.9. Cloning Concatemers and Sequencing

1. Digest 2 μg of pZErO-1 with 10 U of *Sph*I for 30 min at 37 °C.
2. Add 1 μL of CIP enzyme (10 U) and incubate for additional 30 min.
3. Dilute to 100 μL with 1X TE.
4. Extract three times with phenol/chloroform.
5. Precipitate the DNA with 1 μL of 20 mg/mL of glycogen, 10 μL of 3 *M* sodium acetate, pH 5.2, and 250 μL of ethanol.
6. Wash the pellet with 70 % ethanol.
7. Resuspend the pellet in 80 μL of 1X TE (25 ng/μL).
8. Mix 7 μL of concatemers, 1 μL of digested pZErO-1, and 1 μL of 10X ligase T4 DNA buffer.
9. Incubate at 50 °C for 2 min, cool on ice for 2 min, and add 1 μL of T4 DNA ligase (400 U).
10. Incubate at 16 °C overnight.
11. Dilute the ligation mixture to 200 μL with 1X TE.
12. Extract with phenol/chloroform.
13. Precipitate the DNA with 2 μL of 20 mg/mL glycogen, 133 μL of 7.5 *M* ammonium acetate, and 820 μL of ethanol. Incubate in dry ice for 5 min. Spin for 15 min at 4 °C.
14. Wash the pellet four times with 70 % ethanol.
15. Resuspend the pellet in 20 μL of 0.3X TE.
16. Dilute the ligation mixture in double-distilled (dd)H_2O (1:5). Add 1 μL of the diluted DNA to 20 μL of Electromax DH10Bs. Mix gently. Transfer to an ice-cold 1-mm electroporation cuvet.
17. Electroporate with a Biorad gene pulser with following settings: 200 Ω, 25 μF, and 2000 V.
18. Immediately add 1 mL of SOC and shake at 37 °C for 45 min. Plate onto Luria-Bertani (LB) plates containing 50 μg/mL zeocin and incubate for 12 h at 37 °C.
19. Pick colonies, prepare DNA, and perform sequencing reactions.

3.10. Sequence Analysis

1. Use the SAGE2000 v.4.5 program to extract tag sequences from raw sequence data and quantitate the tags as following.
2. Select "New Project" from the "Project" pull-down menu.
3. Change the parameters as follows: "Anchoring Enzyme," *Nla*III–CATG; "Tag Length," 17; "Ditag Length," 36. Click on "Start."

4. A project summary window appears after you click on "Continue."
5. Put all the raw sequence files (*.seq) into the folder where you created a new project.
6. Select "Add Tags" from the "Project" pull-down menu. The sequence files should exist in the "Select Files" field.
7. Check "Auto Analyze" and click on "Analyze."
8. Close the SAGE2000 program by clicking on "Quit" from the "Project" pull-down menu.
9. Open LongSAGE.mdb file with Microsoft Access program (*see* **Note 10**).
10. To create a reference tag sequence library, download the whole genomic DNA sequence from the University of California, Santa Cruz (UCSC) or National Center for Biotechnology Information (NCBI) databases. List forward- and reverse-strand 21-bp tag sequences generated from *Nla*III-cutting sites, and position on the chromosome using a text editor program or simple custom program coded by Perl, C, or C++ language (*see* **Note 11**).
11. Import the reference library to Microsoft Access, identify the tag positions using the "Create Query" option, and map onto the chromosome.
12. Download gene information from the NCBI or UCSC databases and calculate the tags' nearest distance to the gene coding start site for yeast or the transcription start site for human.

4. Notes

1. Be careful not to make foams that reduce sonication efficiency.
2. The average chromatin size should be 300–500 bp, and the approximate DNA concentration is 0.2 mg/mL after removing RNA.
3. This is a preclearing step to reduce nonspecific binding of chromatin to protein A sepharose beads.
4. A control immunoprecipitation with immunoglobulin (IgG) antibody should be carried out in parallel.
5. This step is for filling-in the complementary strand of the linker region.
6. After finishing the PCR amplification, load 3 μL of the reaction mixture onto an agarose gel to examine the DNA. If it is insufficient, then perform three more cycles. Analyze the DNA again.
7. Unlike in the LongSAGE protocol, Linkers 1B and 2B should not be phosphorylated, as they may self-ligate.
8. Test PCR reactions should be performed to optimize the amplification with different dilutions of template (0.1, 0.03, and 0.01 μL of DNA) and different cycle numbers (25, 28, 30 cycles).
9. Check the ligation products under ultraviolet. Most of them are between 400 and 2000 bp.
10. Refer to the Microsoft Access manual for more details.

11. The yeast and human genomic DNA libraries with 21-bp tag sequences containing *Nla*III sites are not available to the public. One can make a tag library from any kind of text editor, including Microsoft Word, by repeating multiple text-search and/or replace. We downloaded yeast and human genomic DNA sequences from the NCBI and UCSC databases, respectively, and used Perl programming language to generate a tag library.

References

1. Felsenfeld, G. and Groudine, M. (2003) Controlling the double helix. *Nature* **421**, 448–453.
2. Kurdistani, S. K. and Grunstein, M. (2003) Histone acetylation and deacetylation in yeast. *Nat Rev Mol Cell Biol* **4**, 276–284.
3. Berger, S. L. (2002) Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.* **12**, 142–148.
4. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
5. Velculescu, V. E., Zhang, L., Zhou, W., et al. (1997) Characterization of the yeast transcriptome. *Cell* **88**, 243–251.
6. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512.
7. Roh, T. Y., Ngau, W. C., Cui, K., Landsman, D., and Zhao, K. (2004) High-resolution genome-wide mapping of histone modifications. *Nat. Biotechnol.* **22**, 1013–1016.
8. Roh, T. Y., Cuddapah, S., and Zhao, K. (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552.

8

5′- and 3′-RACE from LongSAGE Tags

Kåre Lehmann Nielsen

Summary

Serial analysis of gene expression (SAGE) studies often yield numerous tags that cannot be mapped to known gene sequences. Intriguingly, these may represent unknown genes, unknown parts of genes, or transcript variants. In order to elucidate the origin of these tags, 3′- and 5′-rapid amplification of complementary DNA ends (RACE) reactions can be performed using primers identical or complementary to SAGE tags. This way, transcript fragments, or indeed the entire uncharacterized transcript, can be cloned and sequenced.

Key Words: PCR; RACE; gene discovery; tag identification.

1. Introduction

One of the most important features of serial analysis of gene expression (SAGE) and LongSAGE is the ability to detect tags from previously uncharacterized genes (***1***;***2***). Even in this age of high-throughput sequencing, most genomes have not been sequenced to any significant gene coverage, and working in these organisms yields many SAGE tags that cannot be mapped to known transcripts (for an example, *see* **ref.** ***3***). Furthermore, for most organisms where a reasonable amount of sequencing has been carried out, the most comprehensive coverage of transcripts is made up by expressed sequence tag (EST) sequences (e.g., TIGR gene indices at www.tigr.org). These are primarily derived from the 5′ end of complementary DNAs (cDNAs), as a result of difficulties in sequencing through the 3′-poly-A tail of eukaryotic transcripts. Therefore, gene sequence coverage is most complete in the 5′-end

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

of genes. In contrast, SAGE tags are derived from the 3′-most CATG. Consequently, many tags stem from parts of genes not presently in the databases and thus cannot be mapped ***(3)***. In some ways, these unknown tags are the most interesting, because they can aid the annotation of coding regions of genome sequences ***(2)*** or might represent entirely uncharacterized genes. In combination with a high-throughput cloning method, SAGE can work as an efficient gene-discovery engine ***(4)***. An efficient method called GLGI (***5***;***6***) has been published for this purpose, but the method suffers the drawback that only the 3′ end of the transcript is amplified and cloned. Because the CATG, on average, is placed near the 3′-end, most frequently the major part of the coding sequence is located 5′ of the SAGE tag.

Rapid amplification of cDNA ends (RACE) ***(7)*** can be used to amplify both the 5′ and the 3′ end of transcripts, provided a single specific stretch of DNA sequence is known. Indeed, from SAGE tags and, especially, LongSAGE tags, a DNA primer that can be used to amplify the transcript from which the tags are derived can be synthesized ***(8)***. The protocol described here has been developed to minimize the amount of optimization required for each individual LongSAGE tag to facilitate the determination of many unknown SAGE tags in parallel.

2. Materials

2.1. Creating a Suitable Environment for RACE Cloning

1. NaOH from Sigma-Aldrich (St. Louis, MO) is dissolved to 1 *M* in water (*see* **Note 1**). Store in a plastic bottle at room temperature.
2. 70 % EtOH (De danske spritfabrikker, Aalborg, Denmark)
3. New box of pipet tips.

2.2. cDNA Template Preparation

1. 10 μ*M* SMART IV oligonucleotide 5′-AAGCAGTGGTATCAACGCAGAGTGG CCATTACGGCCGGG-3
2. 10 μ*M* CDS III oligonucleotide: 5′-ATTCTAGAGGCCGAGGCGGCCGACATG-d(T)30VN-3′
3. 10 μ*M* 5′-PCR primer: 5′-AAGCAGTGGTATCAACGCAGAGT-3′
4. First strand buffer (5X): 250 m*M* Tris-HCl, 30 m*M* $MgCl_2$, and 375 m*M* KCl. Store at −20 °C.
5. DTT: 20 mM dithiothreitol (DTT) in water. Store at −20 °C.
6. Deoxynucleotides, dATP, dTTP, dGTP, and dCTP (100 m*M* stock solutions) from Fermentas (Burlington, Ontario, Canada). Prepare a mix of all four (final concentration 25 m *M* of each). Store at −20 °C.
7. Taq polymerase 1 U/μL from Fermentas. Store at −20 °C.

8. Taq reaction buffer (10X): 100 m*M* Tris-HCl (pH 8.8), 500 m*M* KCl, 8 % Nonidet P40. $MgSO_4$ (50 m*M*) and $MgCl_2$ (25 m*M*). Chemicals for PCR are conveniently stored together at −20 °C.

2.3. TAE-Agarose Gel Electrophoresis

1. TAE running and gel buffer (50X): 2 *M* Tris, 1 *M* acetic acid, and 50 m*M* ethylenediamine tetraacetic acid (EDTA), pH 7.6–7.8. Store at room temperature.
2. SeaKem GTG agarose (Cambrex, East Rutherford, NJ). Stored dry at room temperature.
3. Gel casting and running apparatus Mini-Sub Cell GT from Biorad (Hercules, CA).
4. EtBr solution: 10 g/L ethidium bromide (Sigma-Aldrich) in water.
5. TAE loading buffer (5X): 0.2 *M* Tris, 0.1 *M* acetic acid, 5 m*M* EDTA in 50 % glycerol. Store at 4 °C.
6. DNA size markers: 1 kb GeneRuler™ (Fermentas). Store at 4 °C.

2.4. Preparation of RACE Template from λ-Phage cDNA Libraries

1. PCI: Molecular biology grade 10 m*M* Tris (pH 8.0), 1 m*M* EDTA-saturated phenol:chloroform:isoamyl alcohol (PCI) (25:24:1) from Sigma-Aldrich. Store at 4 °C.
2. SM buffer: 50 m*M* Tris-HCl, pH 7.5, 100 m*M* sodium chloride, 8 m*M* magnesium sulfate
3. DNase I solution: 1 mg of DNaseI (Sigma-Aldrich) is dissolved in 10 mL 50 m*M* Tris-HCl pH 7.5. Store at −20 °C.
4. 0.5 *M* EDTA (*see* **Note 2**) (AppliChem, Darmstadt, Germany). Store at room temperature.
5. 10 % (w/v) sodium dodecyl sulfate (SDS) (*see* **Note 3**) (AppliChem). Store at room temperature.
6. 2-propanol from Sigma-Aldrich.
7. 100 % and 70 % EtOH, molecular biology-grade, from De Danske Spritfabrikker.
8. TE buffer: 10 m*M* Tris-HCl, 1 m*M* EDTA pH 7.0. Store at room temperature.
9. RNase I solution: 1 mg DNase free RNase I (Sigma-Aldrich) is dissolved in 1 mL 50 m*M* Tris-HCl, pH 7.5. Store at −20 °C.
10. Proteinase K solution: A 20 mg/mL of proteinase K can be obtained from Fermentas. Store at −20 °C.

2.5. The SAGE-RACE Reaction

1. Primers should be reverse-phase cartridge or high-performance liquid chromatography (HPLC)-purified from TAGC (Copenhagen, Denmark). Dissolve the lyophilized primer in water to a final concentration of 100 μ*M* and store at −20 °C. Prior to use, dilute an aliquot to 10 μ*M*. Gene-specific primers should be identical to the LongSAGE tag (including CATG) for 3′-RACE or reverse complementary to the LongSAGE tag (including CATG) for 5′-RACE. The vector-specific primers should

be the T7-primer (TAATACGACTCACTATAGGG) for 5′-RACE and the M13-21 (GTAAAACGACGGCCAG) for 3′-RACE when using λ-phage ZAP cDNA as template and 5′-PCR primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′) for 5′-RACE and 5′-TAGAGGCCGAGGCGGCCGAC-3′ for 3′-RACE.
2. Deoxynucleotides, dATP, dTTP, dGTP, and dCTP (100 m*M* stock solutions from Fermentas). Prepare a mix of all four (25 m*M* each) and use in the experiments. This mix should be stored at −20 °C.
3. Taq polymerase, or a mixture of Taq and a proofreading polymerase should be used. We are using either Taq from Fermentas or Platinum Taq High Fidelity polymerase from Invitrogen (Carlsbad, CA) (*see* **Note 4**).
4. Taq reaction buffers (10X): for the platinum enzyme, the reaction buffer is 600 m*M* Tris sulfate (pH 8.9), 180 m*M* ammonium sulfate. For the Fermentas enzyme, 100 m*M* Tris-HCl (pH 8.8), 500 m*M* KCl, 8 % Nonidet P40.
5. $MgSO_4$ (50 m*M*) and $MgCl_2$ (25 m*M*) should be used with the Invitrogen and Fermentas enzyme, respectively. Store at −20 °C.

2.6. *Cloning of RACE Products into* Escherichia coli

1. TOPO-TA cloning kit from Invitrogen.
2. *Escherichia coli* TOP10 electrocompetent cells (Invitrogen). Store at −80 °C.
3. Autoclaved Luria-Bertani (LB) liquid media, pH 7.2 (10 g/L tryptone, 5 g/L yeast extract, and 10 g/L sodium chloride [Bie og Berntsen, Rødovre, Denmark]). Store in the dark at room temperature.
4. Electroporation cuvets (0.1 cm; BioRAD).
5. Electroporation apparatus, Easyject prima (Equibio, Ashford, UK)
6. LB plates containing 1.2 % agar (Bie og Berntsen) containing 100 μg/mL amplicillin (Sigma-Aldrich). Store for a maximum of 21 d at 4 °C (*see* **Note 5**).

3. Methods

The RACE cloning of SAGE tags is not a simple process for all SAGE tags. Depending on the nature of the SAGE tag (i.e., GC content, secondary structure, etc.) some RACE products are more easily obtained than others, and some reactions fail to provide a specific amplification product. Furthermore, most reactions will contain spurious amplification products, and it is therefore of vital importance that multiple RACE clones are sequenced to ensure that the RACE product is indeed derived from the mRNA from which the SAGE tag originates. Depending on the organism under investigation and the extent of gene sequence coverage of that organism, often a quite substantial number of SAGE tags are derived from unknown transcript sequences. It is therefore desirable to set up the reactions in parallel and avoid optimization of individual reactions. To facilitate the sequencing of RACE clones, the 96-well

microtiter plate format is preferred. For this reason, we work batches of eight RACE reactions and sequence 12 clones of each.

3.1. Creating a Suitable Environment for RACE Cloning

Successful amplification of DNA fragments using a single gene-specific primer and a common primer from a complex solution such as cDNA can be tricky. For rare transcript, it involves the use of many cycles (up to 40) of amplification and is therefore very sensitive to even minute amounts of contaminating DNA. Contrary to what most people think, autoclaving is not always sufficient for decontamination. The DNA molecule is remarkably stable and it has been shown that although most DNA is broken into small, 20- to 30-bp fragments, larger fragments can survive *(9)*. Moreover, tips, tubes, solutions, and even some pipets may be autoclaved, but it is not feasible to autoclave tabletops, etc. To minimize the risk of contamination when performing RACE, we wash the pipets and the tabletop in 1 *M* NaOH followed by water and 70 % ethanol. We always use a new pack of pipet tips and new PCR tubes or plates.

3.2. cDNA Template Preparation

A high quality of poly-A primed cDNA template is fundamental to successful RACE cloning. There are a variety of different kits for preparing cDNA. We are using the BD-SMART cDNA kit from Clontech (Mountain view, CA) to prepare double-stranded cDNA (*see* **Note 6**). This works well in our hands (*see* **Note 7**).

1. Mix 1 µg of DNA with the 1 µL of the SMART IV and 1 µL of the CDSIII oligonucleotides in a final volume of 5 µL. Incubate at 72 °C for 2 min. Place on ice for 2 min and spin briefly to collect condensation to the bottom of the tube.
2. Add 2 µL 5X first strand buffer, 1 µL DTT, 0.5 µL dNTP mix, and 1 µL reverse transcriptase (*see* **Note 8**). Incubate at 42 °C for 1 h (*see* **Note 9**). Place on ice.
3. Transfer 2 µL to a new tube and add 80 µL water, 10 µL of 10X PCR buffer, 1 µL dNTP mix, 2 µL 5′-PCR primer, 2 µL CDS III primer, and 1 U Taq polymerase.
4. Following initial denaturation for 1 min at 95 °C, perform 24 cycles of 95 °C for 15 s and 68 °C for 6 min.
5. Analyze 5 µL on a 1 % TAE-agarose gel (*see* **Fig. 1**).
6. The PCR product can be used as template immediately or stored at −20 °C for later use.

3.3. TAE-Agarose Gel Electrophoresis

1. Mix 0.5 g agarose (for a 1 % gel) with 49 mL of water and 1 mL of 50X TAE. Melt agarose in microwave (*see* **Note 10**) and add 2µl of EtBr.

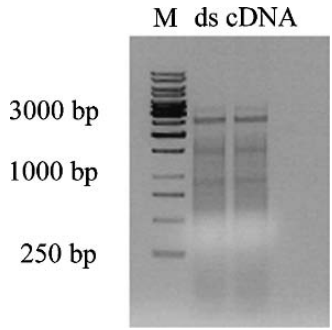


Fig. 1. Double-stranded complementary DNA (cDNA) synthesis product. Lane 1, DNA size marker; lanes 2 and 3, cDNA obtained from potato tuber at 8 wk after planting after 24 cycles of amplification. A smear of DNA is visible up to 3000 bp, with several prominent bands present. These bands represent especially abundant cDNA species.

2. Seal a gel casting form with tape (*see* **Note 11**), pour the melted agarose into the form, and insert a comb. Leave on the bench top until hardened.
3. Remove comb and insert in horizontal gel apparatus and submerge in 1X TAE buffer (*see* **Note 12**).
4. Mix the sample with one-fifth volume of 5X loading buffer and load into wells. Include a suitable DNA size marker.
5. Electrophorese at 10 V/cm until the bromphenol blue is about half the distance of the gel (*see* **Note 13**).
6. Visualize DNA bands by exposure to ultraviolet light.

3.4. Preparation of RACE Template From λ-Phage cDNA Libraries

Preparation of template suitable for RACE can also be made from pre-existing cDNA libraries (*see* **Note 14**).

1. To obtain DNA from λ-phages, add 10 μL of a 100 μg/mL DNAse I to 800 μL of λ-phage stock ($10^5 - 10^6$ pfu) in SM buffer. Incubate at 37 °C for 1 h.
2. Disruption of phage particles and inactivation of DNaseI is achieved by adding 50 μL EDTA solution and 50 μL of SDS solution followed by incubation for 15 min at room temperature.
3. Protein and SDS is removed by extracting with 500 μL of PCI (*see* **Note 15**), mixing vigorously for 30 s, and centrifuging at 10,000 rpm in a benchtop centrifuge. The upper aqueous phase is carefully transferred to a new tube (*see* **Note 16**) and the extraction is performed three times in total.
4. To precipitate DNA, add 600 μL 2-propanol and centrifuge at maximum speed in a benchtop centrifuge for 15 min. Discard the supernatant. A quite large pellet should

be visible. Wash the pellet twice with 70 % ethanol (*see* **Note 17**). Leave the tube on the bench top with the lid open to allow evaporation of residual ethanol for 15–30 min (*see* **Note 18**). Dissolve the pellet in 200 µL of TE (*see* **Note 19**).

5. To remove RNA, add 10 µL of RNase I. Incubate at 37 °C for 30 min. In order to inactivate RNase I and degrade contaminating protein, add 2 µL of proteinase K. Incubate at 37 °C for 30 min. Extract twice with 200 µL PCI as above and precipitate the DNA with 160 µL 2-propanol, and spin in a benchtop centrifuge at 10,000 rpm for 15 min (*see* **Note 20**). Wash with 70 % ethanol and leave the tube with the lid open for 15–30 min.
6. Dissolve the DNA in 150 µL of TE (*see* **Note 19**).
7. Analyze 5 µL on a 1 % TAE-agarose gel (*see* **Fig. 1**).

3.5. The SAGE-RACE Reaction

1. Prepare the following PCR profile on the thermocycler. Following initial denaturation at 94 °C for 2 min, 35 cycles of 94 °C for 30 s, 50 °C for 30 s, and 72 °C for 2 min should be included. Finally, allow all partial extension products to be completed by incubating at 72 °C for 10 min (*see* **Note 21**).
2. In a new PCR tube or plate, add 6.5 µL water, 1 µL template DNA (*see* **Note 22**), and 2.5 µL gene-specific primer. Leave on ice.
3. Prepare a master mix by combining, on ice, 7.5 µL water, 2.5 10X PCR reaction buffer, 2.5 µL common primer, 0.25 µL dNTP mix, 2 µL $MgCl_2$, and 0.5 µL Taq polymerase. Multiply the volumes by the number of samples to be prepared plus one. Remember to include a positive and a negative control (*see* **Note 23**). Add 15 µL of master mix to each sample.
4. Let the thermocycler preheat to 75 °C before inserting the tubes or the plate into the thermocycler (*see* **Note 24**).
5. Analyze 5 µL of each reaction on a 1 % TAE-agarose gel. *See* **Fig. 2** for typical results.

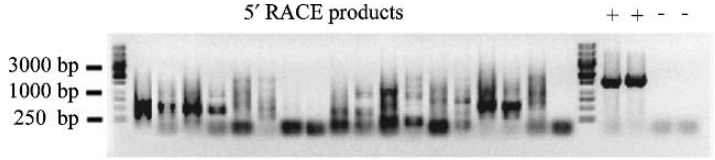


Fig. 2. Typical 5′-rapid amplification of complementary DNA ends (RACE) products from Long serial analysis of gene expression (SAGE) tags. Lanes 1 and 20, DNA size marker; lanes 2–19, RACE products obtained using gene-specific primers complementary to LongSAGE tags and λ-ZAP-potato library DNA as template; lanes 21 and 22, positive control; lanes 23 and 24, negative control.

3.6. Cloning of RACE Products into E. coli

1. Combine 2 μL water with 2 μL RACE product, 1 μL diluted salt solution and 1 μL Topoisomerase-vector complex (supplied with the kit). Leave on the bench for 5–30 minutes and place on ice. The reactions should be used for bacterial transformation the same day.
2. Thaw 50 μL of electrocompetent *E. coli* (*see* **Note 25**), and transfer 2 μL of the TOPO reaction above to the cells. Mix by gently by stirring with the pipet tip a few times. No incubation time is necessary.
3. Gently transfer the reaction to an electroporation cuvet, insert into the electroporator and pulse at 1800 V/cm (*see* **Note 26**). Immediately add 250 μL of SOC medium (*see* **Note 27**).
4. Transfer to a 10-mL culture tube and incubate for 1 h.
5. Spread 25 and 100 μL on two ampicillin containing agar-plates (*see* **Note 28**).
6. Incubate overnight at 37 °C, and pick 12 colonies of each for sequencing.

4. Notes

1. Unless stated otherwise, all solutions should be prepared in water that has a resistivity of 18 MΩ/cm and an organic content of less than five parts per billion. However, for small volume solutions up to 10 mL, molecular biology-grade water from AppliChem can be used to ensure reproducibility even when the local water systems fail.
2. Sodium-EDTA cannot be dissolved to 0.5 *M* because it is an acid, and EDTA is less soluble in acidic solutions. Complete dissolution can be achieved by adding small amounts of sodium hydroxide during dissolution. Care should be taken that pH does not increase above 8.0.
3. SDS foams very easily, and the easiest way to make a solution is to add the measured amount of powder to half the final volume of water in a measuring cylinder and then carefully add the rest of the water. Close the cylinder with Parafilm (thoroughly) and gently invert the tube until the SDS is dissolved.
4. We have not observed great and consistent differences between different polymerases used.
5. A large portion of agar containing LB media without antibiotics can be divided into aliquots of 350 mL in 500-mL bottles and autoclaved. These bottles can be stored in the dark at room temperature. When needed, they can be melted in a microwave oven and allowed to cool to about 45 °C before addition of a sterile-filtered solution of an antibiotics stock solution.
6. Double-stranded cDNA can be used for both 5′ and 3′-RACE. However, for 3′-RACE reactions, first strand cDNA synthesis is sufficient.
7. SMART cDNA synthesis relies on the template-switching ability of Moloney murine leukemia virus reverse transcriptase. Typically, an RNase H^- derivative (such as PowerScript from Clontech) is used to enrich for full-length clones.

Deletion mutant variants such as SuperScript™ III (Invitrogen) do not exhibit sufficient template-switching activity.

8. Do not vortex the solution; intracellular enzymes are easily inactivated by oxidation of cysteine residues. Instead, mix gently by pipetting up and down five times, avoiding bubbles.
9. It is important to use an air incubator or a PCR machine with a heated lid. Otherwise, condensation will appear on the top of the lid. Because the reaction volume is only 10 μL, the concentration of nonvolatile molecules will change significantly during incubation and cDNA synthesis may fail.
10. Use a 100-mL bottle for 50 mL gel to avoid boiling over. Put the lid on loosely to minimize loss by evaporation.
11. Not all types of tape will stay sealed. Autoclave tape works fine and is often at hand, but is the most expensive. Cheaper alternatives can be found.
12. Add just enough buffer to submerge the gel. Too much buffer will only increase electrophoresis time and buffer consumption.
13. The bromphenol blue dye migrates as approx 400 bp in a 1 % TAE-agarose gel.
14. We have performed most of our reactions on the λ-ZAP-potato library described in **ref.** ***(10)***. In our hands, this works better than cDNA, especially for the 5′-RACE reactions.
15. Remember that it is the organic phase (the lower) that should be added to the sample. The upper phase consists of excess buffer used for the saturation and pH adjustment of the organic phase.
16. It is crucial for the final result that none of the inter phase or the organic phase is transferred to the new tube. Ten percent of the aqueous phase should be discarded in each extraction.
17. For convenience, we store solutions of 70 % ethanol at 4 °C because the pellet tends to slip more easily when washed with room-temperature ethanol than with cold.
18. Do not use a dessicator. λ-DNA is 44-kb, and can be very hard to dissolve if dried completely.
19. The pellet is often hard to dissolve and it may require incubation at room temperature for an hour. Do not heat the sample, because incomplete annealing of denatured λ-DNA will cause the solution to be viscous.
20. The pellet is significantly smaller than before because the RNA has been removed. It can sometimes be difficult to see, as it is often smeared over a large area at the backside of the tube.
21. The annealing temperature is, of course, primer-specific, but 50 °C seems to be a good compromise for LongSAGE tags. If a reaction fails, we usually try it at 45 °C and 55 °C. It seems to be more difficult to tune the fidelity by altering the Mg^{2+} concentration. This is presumably because, when we have one primer that is common to all or nearly all DNA fragments present, very high specificity of the gene specific primer is important in order to avoid any nonspecific amplification products.

22. When using cDNA, we dilute it 10 times. When DNA purified from λ-phages is used, we dilute it 25–50 times. Approximately 10 ng of template DNA is desirable.
23. We use a plasmid containing the common primer site and an insert for which we have a primer as the positive control, and a reaction excluding the gene-specific primer as the negative control.
24. This is not necessary if an antibody inhibited polymerase (Hot-start polymerase) is used.
25. We use the *E. coli* Top10 cells, which work fine, but other *E. coli* cells may be used. Be careful when thawing cells; they are very vulnerable to lysis by osmotic shock, because they are in an almost nonsalt buffer. Excessive heating (room temperature), vortexing, or pipetting up and down will decrease transformation efficiency dramatically. Either thaw quickly between the fingers (be careful) or slowly on ice.
26. The exact settings of the electroporator are dependent on the size of the gap between the electrodes in the cuvet and the electroporator apparatus. Consult the manual for the apparatus.
27. We are using SOC medium because slightly more transformants are obtained this way. However, approximately half the transformants are obtained using LB. Usually, this is sufficient.
28. It can be difficult to spread 25 μL evenly, so we usually add 50 μL of LB media to the plate before adding the cells, and then spread. We find spreading bacteria using approximately five small, sterile, disposable glass beads more convenient than using Drigalski spatulas.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
2. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512.
3. Nielsen, K. L., Grønkjær, K., Welinder,K. G., and Emmersen,J. (2005) Global transcript profiling of potato tuber using LongSAGE. *Plant Biotechnol. J.* **3**, 175–185.
4. Boheler, K. R. and Stern, M. D. (2003) The new role of SAGE in gene discovery. *Trends Biotechnol.* **21**, 55–57.
5. Chen, J. J., Rowley, J. D., and Wang, S. M. (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *Proc. Natl. Acad. Sci. USA* **97**, 349–353.
6. Chen, J., Lee, S., Zhou, G., and Wang, S. M. (2002) High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3′ complementary DNAs. *Genes Chromosomes Cancer* **33**, 252–261.

7. Frohman, M. A., Dush, M. K., and Martin, G. R. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85**, 8998–9002.
8. van den Berg, A., van der Leij, J. and Poppema, S. (1999) Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. *Nucleic Acids Res.* **27**, e17.
9. Elhafi, G., Naylor, C. J., Savage, C. E., and Jones, R. C. (2004) Microwave or autoclave treatments destroy the infectivity of infectious bronchitis virus and avian pneumovirus but allow detection by reverse transcriptase-polymerase chain reaction. *Avian Pathol.* **33**, 303–306.
10. Crookshanks, M., Emmersen, J., Welinder, K. G., and Nielsen, K. L. (2001) The potato tuber transcriptome: analysis of 6077 expressed sequence tags. *FEBS Lett.* **506**, 123–126.

2

Tag Extraction and Analysis

9

Extraction and Annotation of SAGE Tags Using Sequence Quality Values

Jeppe Emmersen

Summary

Data analysis of serial analysis of gene expression (SAGE) tag experiments begins with the extraction of tags from single-pass sequence files of ditag concatemers. When using DNA base quality values generated during base calling, it is possible to control the false-positive discovery rate of unique tags. This chapter describes how to set up a system for generating tag lists from quality associated sequence data.

Key Words: Phred; quality values; mapping; sequence databases; UniGene.

1. Introduction

The serial analysis of gene expression (SAGE) method for transcriptome profiling has been termed a "digital" method compared to the "analog" method of microarrays. The term "digital" implies that SAGE data analysis is simply a matter of counting the sequence tags after sequencing a number of concatemers. In reality, the SAGE data only becomes digital after tag extraction. To obtain a reliable, digital tag count using SAGE analysis, one must address sequence quality variation and variations in ditag length during the extraction process and, after tag extraction, tag ambiguity during tag mapping. Here, we describe some of the problems associated with these two issues in more detail and discuss how they can be minimized. The Methods section describes how to make use of sequence quality data for tag extraction using a simple Perl script and how to map the resulting tag list to a fasta-formatted sequence database.

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

1.1. Variations of the MmeI Site Preference

In LongSAGE, the type I restriction enzyme *Mme*I combined with type II enzymes is used to generate the SAGE tags. *Mme*I cuts the DNA template 21–22 bases downstream from the recognition site, generating a 2-base overhang at the 3′ end ***(1)***. When a ditag is formed, only tags with complementary overhangs can ligate together. Thus, when extracting ditags, the expected size range of ditags is 36 bases to 38 bases (equal to two times 21 or 22 minus 2-base overlap, minus the second anchoring enzyme recognition sequence; CATG, for *Nla*III). The distribution of ditag lengths of a human pancreas SAGE library is seen in **Table 1** ***(2)***. Notice the near 1:2:1 ratio of ditags with lengths 36, 37, and 38. Ditags with length 37 are due to a combination of one short tag and one long tag.

When extracting ditags from the raw sequence, the search algorithm finds the CATG punctuation generated by the anchoring enzyme recognition sequence. Thus, in the case of *Nla*III, each ditag can be extracted by searching for the pattern −CATG − (34–36 bases from the complementary DNA [cDNA]) − CATG-.

Sometimes ditags of larger sizes can be observed. These are probably artifacts of the ligation process, whereby hybridization occurs between tags overlapping on one base (ditag size of 39 bases) only or even blunt-ended hybridization (ditag size of 40) (*see* **Table 1**). Such anomalies are to be expected, given the large amount of different molecules interacting during ditag generation. Although the larger ditags, in theory, gives more sequence information, which is important for tag mapping, tag extraction is performed using the lowest allowable tag size to ensure the most correct tag count. For LongSAGE, the size is usually set to 17 bases, and any additional sequence information is discarded.

Table 1
Distribution of Ditag Lengths

Ditag length	No. of ditags
39	115
38	4288
37	12262
36	6540

From the same sequence data used for **Fig. 1,** the distribution of ditag lengths was calculated. A Phred quality threshold of 20 and tag length of 17 bases was used.

1.2. Other Anchoring Enzymes

The punctuation pattern generated from the anchoring enzyme recognition site sequence is the fundamental principle behind ditag extraction. This means that it is easy to change the parameters of the ditag extraction, for example to use another anchoring enzyme such as *Sau*3A (GATC) or a different type II restriction enzyme.

A reason to use more than one anchoring enzyme would be to minimize the number of genes not containing the recognition site of a particular anchoring enzyme. For the *Nla*III enzyme, the most widely used anchoring enzyme, it has been shown that approx 2 % of gene sequences in *Drosophila melanogaster* do not contain the recognition sequence CATG ***(3)***. Similar levels of tag ambiguity have been found in other organisms ***(3,4)***.

1.3. Sequence Quality

Most unique SAGE tags found after ditag processing are found only once (*see* **Fig. 1**). These tags are due either to sequencing errors, which are most prominent at the 5′ and 3′ ends of a sequence ***(5)***, or to the use of single-pass sequencing of SAGE concatemers. Ideally, each concatemer would be sequenced twice to confirm each base call. However, this would decrease throughput by 50 %; thus, sensitivity is preferred over specificity for normal SAGE sequencing. Specificity, however, can be increased without such a dramatic drop in efficiency by using a basecaller such as Phred, which attaches a quality value with each base determined ***(6)***. The Phred quality value is a log-transformed measure of how certain each basecall is. The direct relationship between quality values and probability of base errors is found by the formula $QV = -Log_{10}(Pe)^{*}10$; Pe is the base error probability. Thus, a quality value of 20 corresponds to a probability of 99 % for a correct basecall and quality value of 10 to a probability of 90 %. When extracting tags, the Phred quality values can be used to control the false discovery rate of SAGE tags. The effect of employing a minimum quality value during tag extraction is shown in **Fig. 1**.

1.4. Annotation of Tags (Mapping) and Search Strategies

After tag extraction, it is necessary to match each tag to a gene, a process called mapping. Mapping is simply a process of matching the tag sequence to a specific gene sequence. This gene sequence should preferably come from a collection of completely annotated mRNA sequences containing the complete mRNA molecule or at least information on the 3′ completeness. This means that choosing the database with which to map tag sequences to genes is an

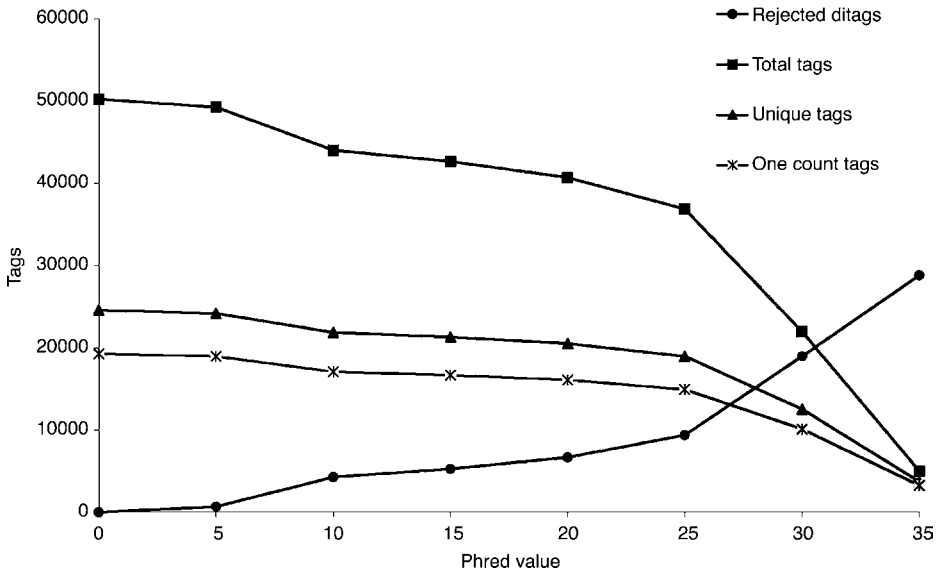


Fig. 1. Library-wide statistics for a human serial analysis of gene expression library. The most important library descriptors are plotted against Phred quality values ranging from zero to 35: Rejected ditags, Total tags, Unique tags, and Single tags.

essential part of tag annotation. Two features on the DNA sequence help us identify the correct gene: the tag sequence itself and the use of *only* the most 3′ anchoring enzyme restriction site of the mRNA sequence.

This suggests that mapping tags to genes is simply a matter of finding the most 3′ sequence corresponding to each tag sequence. Nevertheless, this task is hampered by a number of problems. First, not all 3′ sequences are unique. Particularly closely related genes may share the same 3′ most sequence. Second, it is not always possible to obtain the complete mRNA sequence. In particular, many mRNA sequences available from public data repositories lack information on both 5′ and 3′ untranslated regions (UTRs) as well as polyA-tail information. This is true even in the advent of large-scale, full-length cDNA sequencing efforts such as the Mammalian Gene Collection or the similar RIKEN *Arabidopsis* initiative ***(7,8)***. One reason is that the main efforts are not aimed at obtaining the entire mRNA sequence, but rather at finding the full-length open reading frame (FL-ORF) of the mRNA.

If a tag does not match any 3′ CATG virtual tags, a less stringent match criterion may be used. One possible method of less-stringent tag matching is blastn, which will find all related sequences. On the homepage of the National Center

for Biotechnology Information (www.ncbi.nlm.nih.gov), a special version of blastn—"Search for short, nearly exact matches"—is available and can be used for tag identification. Care should be taken when using this approach, as the number of possible matches will increase dramatically. Nevertheless, in the case of alternative splicing transcripts or sequence polymorphisms, this is probably the best method, as the BLAST web interface gives direct access to the underlying gene models. Each case must be reviewed very carefully, so this method is probably best employed for tags of high interest, i.e., tags found to be differentially expressed between different tissues. Additional methods such as SAGE- rapid amplification of cDNA ends (RACE) can then be used to verify the tag annotation ***(9)***. There have been reports of the finding of antisense tags corresponding to antisense RNA ***(10)***. Such tags may be found by generating virtual tags from all possible CATG sites (*Nla*III) with each gene. As the number of false-positives will increase by such a search, results must be confirmed by other methods such as SAGE-RACE or Northern blots ***(10)***.

1.5. There is no Large-Scale mRNA Database for my Organism!

For most organisms, there are no large databases of mRNA models available. In this case, it is possible to use assemblies of expressed sequence tag (EST) sequences for tag mapping (Unigenes). A number of resources publish large-scale EST assemblies for download ***(11,12)***. As the orientation of an EST assembly is not guaranteed, mapping of tags might include both the most 3′ and the most 5′ tag site.

If no sequence resource can be found in the public databases, the only solution is to obtain EST sequences from the dbEST subdivision of GenBank or to construct a new EST library. The latter approach is of course the most cumbersome and expensive, but it has the added benefit that the sequences originate from the exact same variety of the organism as used for SAGE analysis. This way, the influence of sequence polymorphisms among varieties of the same species is minimized. Subsequent mapping of tags to the custom database may then improve compared to using a generic database for one organism. An example of this approach is shown in the SAGE analysis of potato tubers from the Kuras cultivar, where both a generic Unigene sequence database consisting of all potato varieties generated by The Institute of Genomic Research (TIGR) and a Kuras tuber-specific EST Unigene database were used ***(10,13)***.

The TIGR-generated database (StGI, Release 9) consisted of 32,553 unigenes, whereas the kuras0-specific database consisted of 1088 unigenes. Mapping of 58,322 Kuras LongSAGE tags found seven tags unique to the Kuras variety among the 50 most abundant tags.

2. Materials

1. A computer with the Perl language installed. For Windows, Perl can be obtained from www.activestate.com free of charge.
2. For base calling DNA sequences with associated quality values, a copy of the base calling program Phred is needed. Phred is free of charge for noncommercial use and can be obtained at www.phrap.org. A license agreement must be completed before Phred is made available.
3. The Perl scripts for tag extraction (sage-phred.pl) and tag mapping (sage-map.pl) can be obtained from je@bio.aau.dk.

3. Methods

The following describes how to set up an environment for importing raw DNA sequences from a MegaBace DNA sequencer (GE Healthcare) and convert these to lists of tag counts using simple Perl scripts. This section assumes the user has limited experience with Linux but has access to a Linux system. Commands to executed are shown in **bold** on separate lines, options for commands are shown in ***bold italics.***

3.1. Setting Up Software

Phred is part of the Phred-Phrap package, written by Phil Green. A commercial version for Microsoft Windows can be obtained from CodonCode (www.codoncode.com).

Assuming the source codes for Linux have been obtained, a binary executable of Phred must be compiled from the source code.

1. The Phred source files come in a zipped tar archive. Create a new folder for Phred: **mkdir Phred**.
2. Move the tar archive to this folder and change the working directory to Phred: **cd *Phred***.
3. Unpack the source files: **tar *–zxf name-of-tar-archive***.
4. Make the executables: **make**.
5. There are now two important files present in the folder: the Phred executable and the data file containing base calling data from different sequencers. The Phred program must be made executable by the system: **chmod *777 phred***.
6. Move the executable to a folder in your path. To see what is in your path, use the env command and look for the Path variable: **env**
7. Move the Phred program and the parameter file phredpar.dat to the folder /usr/bin/ (root privileges may be needed for this):
   mv phred /usr/bin/.
   mv phredpar.dat /usr.bin/.

8. Now the Phred program must be made aware of the path to the parameterfile. This is done by adding the following line to the bash shell initialization file /etc/profile using a text editor: **PHRED_PARAMETER_FILE = "/usr/bin/phredpar.dat" export PHRED_PARAMETER_FILE**

3.2. Base Calling With Phred

The system is now set up for base calling with Phred. Copy the sequencer files from the sequencing instrument to a separate folder in the home directory. Here we will name this folder *Megabace_in*. For the MegaBace sequencer, these files end in *.esd*. Each file should have a unique name, otherwise they will be over written during the base calling process. Make a folder for the Phred files: **mkdir *Megabace_phred***

Base calling is executed by the following command:

Phred *–id Megabace_in –pd Megabace_phred*

3.3. Tag Extraction

Tag extraction can now be performed using the script sage-phred.pl. To obtain an overview of options for tag extraction, execute sage-phred.pl without options as follows:

perl *sage-phred.pl*

usage: sage-phred.pl [options] <Phred sequence directory>

options

-q	*<quality threshold>*	*[20]*
-d	*<duplicates included>*	*[default − off = 0; on = 1]*
-l	*<length of tag>*	*[default = 17]*
-k	*<keep good tag from bad ditags>*	*[default − off = 0; on = 1)*

Options for sage-phred.pl are always numeric, so to include duplicate ditags in the tag counts, type **sage-phred.pl –d 1**.

The most important parameter is the -q option. This is set to a default of 20 but can be changed to more or less stringent parameters. The length of the SAGE tag extracted is set by the -l option. By setting this parameter to 10, it is possible to generate tag lists that can be compared to tag lists from the original protocol. The -k option set to 1 makes the script keep tags consisting of bases with quality values of at least those set as the threshold, even though the partner tag of a ditag contains bases below this threshold. With this in mind, it is now possible to make a tag extraction from the sequence folder generated previously: **sage-phred.pl *–q 20 –l 17 –k 1 –d 0 Megabace_phred***

will extract all tags from the folder Megabace_ phred, using a quality threshold of 20, extract tags of length 17 (excluding anchoring enzyme site), include all good tags, and discard duplicate ditags.

The folder **Megabace_phred–results–20–17–0** contains the results from the extraction; the numbers in the folder name indicates the most important parameters of the extraction.

Five files are generated in the tag extraction process:

1. A tag count file consisting of tags separated by a tab.
2. A tag file in fasta format, with the header defined by the tag sequence and the tag count. This file is provided for easy mapping of nonexact tags with blastn.
3. A list of duplicated ditags.
4. A summary of tag extraction, listing the unique tags, total number of tags, ditags rejected from bad sequence quality, and distribution of ditag lengths.
5. A ditag file.

3.4. Mapping Tags to a Sequence Collection

1. The SAGE tag list generated by sage-phred.pl can be used to map the tags to a fasta-formatted sequence collection using the sagemap.pl script. Before using this script, line 22 must be edited to reflect the location of the databases: **my $databasepath = "path to fasta formatted databases"**.
2. Executing sagemap.pl with the option **-d view** will provide a list of available databases in the database folder.
3. Mapping the tags extracted from Megabace_phred folder is performed the following way:
   cd ***Results-Megabace_phred-20-17-0***
   sagemap.pl ***–t tags-Megabace_phred-20-17-0.tsv –c databasefile***
4. This will produce two files in the results directory:
   Annotation-Megabace_phred and **SAGEmap-Megabace_phred** (*see* **Note 1**).
   The **Annotation-** file contains all mappings in tabulated columns. The **Function-** file contains all annotation extracted from the sequence header. These files contain multiple mapping concatenated in one column.
5. The **SAGEmap-** file contains only one mapping in tabulated columns. If the sagemap.pl script is invoked using the parameter **-a 1**, all possible virtual tags will be extracted. The complete Sense tags mappings are found the SAGEmap-Sense file and antisense mappings in the **SAGEmap-Antisense-** file. Using this option will increase the time needed for computation considerably, as every possible tag-virtual_tag combination is searched. The mapping types can be referred to as "genome mode" vs "cDNA" mode.
6. When mapping tags to sequence databases, the number of identified tags can be increased by using different databases or by using non-exact matching algorithms such as BLAST. If each database mapping is ranked according to fidelity, this

approach may yield additional results. For instance, mapping human sequence databases can be ranked from RefSeq, available from GenBank; Predicted cDNA, available from complete genome assemblies at EnSembl; or cDNA Unigene collections, available from either TIGR or Unigene (*see* **Note 2**).

4. Notes

1. Tabulator separated files are easily imported into any spreadsheet program, such as Excel, for further analysis.
2. As a final resort, blastn can be used to annotate important tags by searching internal tag sequences. This makes the annotation more uncertain, but may help identify alternative splice variants.

References

1. Daisuke, S., Sugao, K., Namiki, S., Tanabe, M., Iino, M., and Hirose K. (2004) Enzymatic production of RNAi libraries from cDNAs. *Nat. Genet.* **36(2),** 190–196.
2. Heidenblut, A. M., Leuttges, J., Buchholz, M., et al. (2004). aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Res.* **32(16),** e131.
3. Pleasance, E. D., Marra, M. A., and Jones, S. J. (2003) Assessment of SAGE in transcript identification. *Genome Res.* **13(6A),** 1203–1215.
4. Nielsen, K. L., Emmersen, J. E., and Welinder, K. G. (2004) Digital transcriptomics—a flavour of SAGE. *Biochemist* **26**.
5. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.* **18**, 3657–3665.
6. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8(3),** 186–194.
7. MGC Project Team. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* **14,** 2121–2127.
8. Seki, M., Narusaka, M., Kamiya, A., et al. (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science* **296,** 141–145.
9. Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M., and Wirth, D. F. (2001) Serial analysis of gene expression in Plasmodium falciparum reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Mol. Biol. Cell. **12(10),** 3114–3125.
10. Nielsen, K. L, Grønkjær, K., Welinder, K. G., and Emmersen, J. (2005) Global transcript profiling of potato tuber using LongSAGE. *Plant Biotech. J.* **3(2),** 175–185.

11. Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28(1),** 141–145.
12. Wheeler, D. L., Church, D. M., Federhen, S., et al. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31(1),** 28–33.
13. Crookshanks, M., Emmersen, J., Welinder, K. G., and Nielsen, K. L. (2001) The potato tuber transcriptome: analysis of 6077 expressed sequence tags. *FEBS Lett.* **506**, 123–126.

10

Correction of Technology-Related Artifacts in Serial Analysis of Gene Expression

Viatcheslav R. Akmaev

Summary

Serial analysis of gene expression (SAGE) is a powerful technique for measuring global gene expression through sampling of transcript tags. SAGE tag collections or libraries serve as a rich data source for differential gene expression analysis, transcriptome mapping, and gene discovery. Transcriptome mapping and gene discovery are facilitated by extensions of SAGE, e.g., Long SAGE, where the transcript tags are elongated by utilization of a different tagging enzyme. SAGE, as a sequencing-based technique, is prone to errors resulting in artifact SAGE tag sequences and erroneous tag numbers. A methodology to pinpoint and correct tag artifacts is necessary to fully exploit the value of large SAGE libraries.

SAGEScreen is a tag sequence correction algorithm. The algorithm is a multistep procedure that addresses error rates and performs ditag and tag processing. The error rate estimates are based on a stochastic model of PCR and sequencing related mutations. The ditag processing step is essential for calculation of unbiased tag numbers, and the tag processing step allows for filtration of tag sequence artifacts and adjustment of tag numbers.

Key Words: SAGE; Long SAGE; PCR errors; sequencing errors; SAGE tag artifacts.

1. Introduction

Serial analysis of gene expression (SAGE) is an abundant source of genomic data. SAGE is a protocol for systematic, high-throughput generation of short expressed sequence tags (ESTs) from a tissue sample, producing a global profile of gene expression. The SAGE technique allows for collection of short mRNA sequence tags from a specific position in the transcript. The tag

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

position is defined by the location of the 3′-most anchoring enzyme restriction site. The most commonly used enzyme for this purpose is *Nla*III. Complementary DNA (cDNA) fragments from cleavage with anchoring enzyme are further processed with a tagging enzyme, typically, a type IIS restriction endonuclease, BsmFI. Following amplification, cloning, and sequencing, the SAGE protocol results in a set of vector inserts from which ditags and, ultimately, tag sequences are extracted and counted. In theory, tags of this length are sufficiently specific to map the transcriptome. In fact, a majority of the human SAGE tags map uniquely to UniGene clusters ***(1)***. Given such a bi-directional map, expression levels of the transcripts are inferred from observations of the SAGE tags. Several groups have advanced the original SAGE protocol with the use of different tagging enzymes ***(2,3)***. Particularly in LongSAGE, a type IIS restriction endonuclease, *Mme*I, cleaves 21–22 bases downstream of the anchoring enzyme restriction site, thus, yielding longer tag sequences. The LongSAGE protocol improves the specificity of the tag-to-transcript mapping, and allows for direct tag sequence queries in the genome.

By nature, the SAGE protocol is subject to sequence errors introduced by PCR and sequencing. Suboptimal fidelity of these procedures can introduce artifact tag sequences. Such "mutations" usually do not occur frequently for individual transcripts, and may have little effect on the quantification of differential expression of moderately expressed genes. Their consequences are greater for expression measurements of rare transcripts and prediction of novel genes. SAGE sequence errors have been extensively studied ***(4)***. In LongSAGE, the expected proportion of tag sequences affected by PCR mutations in a tag library is about 3.5% and the expected proportion of tags affected by sequencing errors is about 15.6%. In combination, these numbers lead to the overall error rate of about 17%. Empirical error rates observed in several LongSAGE libraries varied from 10% to 19% ***(4)***. The volume of errors in SAGE is worth attention. Approximately 80% of the artifact tags are one-base variants of authentic tag sequences, and more than 95% of the mutants are one- or two-base variants. Erroneous tags usually form clusters of sequences related to particular genuine tag sequences. The hierarchical structure of tag sequences in SAGE libraries is exploited in *SAGEScreen*.

2. Materials

1. DNA Sequencing Analysis software (Applied Biosystems, Foster City, CA).
2. Phred software (Laboratory of Phil Green. Genome Sciences Department, University of Washington).
3. Programming language, e.g., C, C++, Perl.

3. Methods

From the theoretical calculations and observed error rates, it is evident that the volume of erroneous SAGE tags is substantial. Artificial tags are rare and diverse, and a majority of them have observations in the single digit numbers. The error rate projections indicate that mutant tag sequences constitute approximately one-sixth of a tag library. On the scale of tag sequences, the proportion of mutant sequences may very well reach 30–40 %. The correction algorithm that we have developed is a multistep approach that exploits the intrinsic structure of SAGE data, utilizes empirical estimates of the error rates, and preserves the discrete and stochastic organization of SAGE tag sampling. The first step of the algorithm is analysis of ditags. Observed duplicate ditags are generally discarded in SAGE data processing to control for PCR amplification bias (*see* **Note 1**). The algorithm finds potentially mutant ditags and removes them from the library with little risk of affecting true tag counts. The next step of the algorithm is the calculation of the empirical error rates from the expression patterns of abundant SAGE tags. Based on these rates, the algorithm seeks clusters of tag variants related to one true tag sequence. Subsequently, the variants are deleted from the data set, and their counts are added to the parent tag counts. At the end, only potentially genuine variant tags are retained in the library. Throughout this chapter, the following notation is assumed: L is the SAGE tag length.

3.1. Ditag Processing (see Fig. 1)

Processing of SAGE ditags requires a tab-delimited list of ditags, with ditag sequences in column one, the ditag counts in column two, and in columns three and four, information about the fidelity of the left and right tag, respectively, based on a sequence data processing algorithm such as *Phred* ***(5)***. The fidelity of the tags is denoted as 1 for being legitimate and 0 otherwise. In situations where fidelity information is not available, for example when processing sequence data with the *DNA Sequencing Analysis* software from **Applied Biosystems,** columns three and four are filled with 1s if the tag sequences are complete and 0s if the sequences contain ambiguous nucleotide symbols such as “N,” “Y,” etc.

Based on the list of ditags, a first-pass SAGE library is created. The first L nucleotides of the ditag sequence are saved as the left SAGE tag and the reverse-compliment last nucleotides of the ditag sequence are saved as the right tag. Tag counts are accumulated for legitimate tags. The tag counts of the low-quality tag sequences are set to zero. Every pair-wise combination of ditag sequences from the ditag list is analyzed.

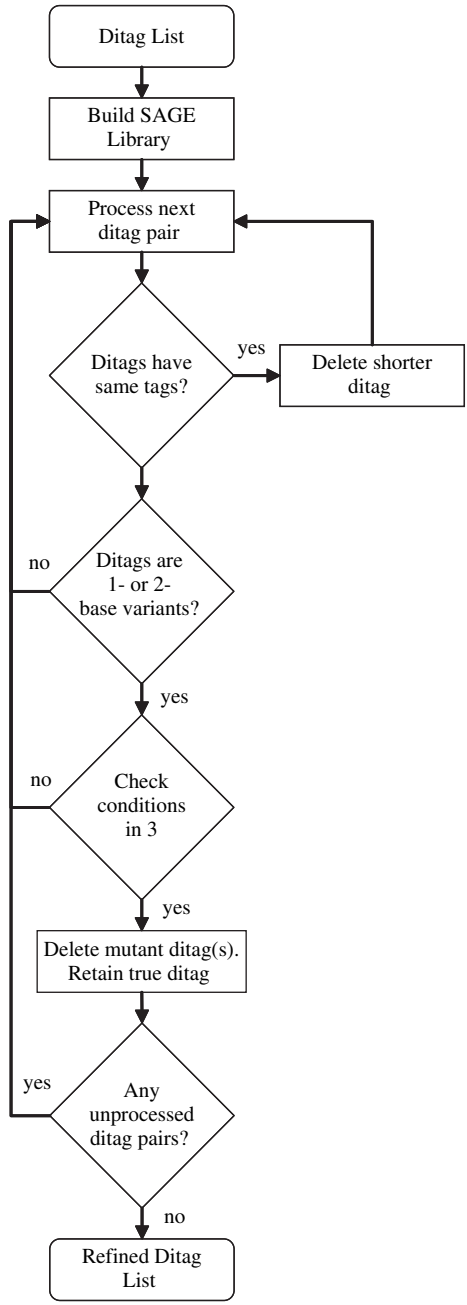


Fig. 1. Ditag processing workflow.

1. Tag combinations are compared. If the two ditags generate identical tag pairs, the shorter ditag is deleted from the list and the next ditag pair is processed.
2. If the two ditags do not generate identical tag pairs, the algorithm checks whether the two ditag sequences are one- or two-nucleotide variants (*see* **Note 2**). If not, the next ditag pair is processed.
3. If the two ditag sequences are one- or two-nucleotide variants, the counts of the four tags are compared. If one of the ditags has the left tag count larger than or equal to the left tag count of the other ditag and the right tag count larger than or equal to the right tag count of the other ditag with one of the tag counts strictly larger, then the ditag with the larger tag counts is retained in the list and the other ditag is deleted. If the left tag with larger count and the right tag with larger count belong to the different ditags, then a new ditag is formed of these two tags and added to the list and the two original ditags are discarded. In all other cases, the both ditags are retained and the next pair is processed.

When all ditag pairs are analyzed, a second-pass SAGE library is built similarly to the first pass library.

*3.2. Estimation of Error Rates (***see** *Fig. 2)*

The algorithm analyzes a tab-delimited list of SAGE tags and their counts. The error rates are estimated from observed mutations in abundant transcripts. The abundance of these transcripts must be large to guarantee reasonable expectations of substitutions, deletions, and insertions. It is recommended that the threshold be set at 50 tag counts for LongSAGE libraries.

1. Abundant tags are selected in the SAGE library, e.g., SAGE tags with counts of 50 and above.
2. For each abundant, or in other words, parent tag, all one- or two-nucleotide variant SAGE tags with lower tag counts are found, and the error-free parent tag count is determined by adding the counts of the variant tags to the parent tag count. The variant tag sequences are sorted into three bins: substitutions, deletions, and insertions. If the mutation type is ambiguously defined, the substitution takes precedence. The counts of the tags in each bin are summed and divided by the error-free parent tag count, thus converting the mutation counts to the mutation frequencies.
3. At this step, the algorithm works with three collections of numbers—the substitution, deletion, and insertion frequencies derived from the parent tags. If the collection size is sufficient (the recommended size of the set is 10 and above), the median of the collection is taken as the mutation frequency estimate, $\hat{f}$.
4. To obtain error rate estimates, the following equation is numerically solved for $\hat{r}$ for each of the three mutant types (*see* **Note 3**):

$$L \cdot \hat{r} \cdot (1 - \hat{r})^{L-1} = \hat{f} \tag{1}$$

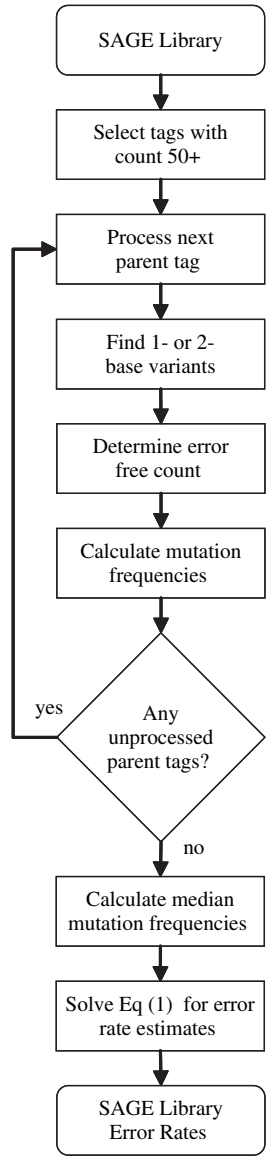


Fig. 2. Workflow for the error rate estimation algorithm.

5. If the size of one of the frequency collections is not sufficient and the mutation frequency is unreliable, it is recommended that the error rates observed in a high-quality, large SAGE library are used. We suggest the following estimates (*see* **Note 4**):

$$\begin{aligned} \hat{r}_{substitution} &= 0.004 \\ \hat{r}_{deletion} &= 0.0006 \\ \hat{r}_{insertion} &= 0.0005 \end{aligned} \tag{2}$$

3.3. Tag Processing (see Fig. 3)

The tag processing step works with two inputs—a tab-delimited SAGE library and error rates. Additionally, the user needs to specify a *p*-value threshold. This threshold is used in the statistical test.

1. Calculation of the parent tag count cutoff:

$$C_0 = \frac{\log(1 - p_0)}{1 - (1 - \hat{r}_{substitution} - \hat{r}_{deletion} - \hat{r}_{insertion})^L} \tag{3}$$

   where p_0 is the *p*-value threshold. The tags with the count larger than the count cutoff, C_0, are selected as parent tags.
2. For each parent tag:
   a. The one- or two-base variant SAGE tag sequences with counts smaller than the parent tag count are located and saved in a cluster.
   b. The tag's adjusted *p*-value is calculated for each tag in the cluster (*see* **Note 5**):

$$p_t = \sum_{i=C_t}^{i=S} \frac{e^{-P_m} \cdot P_m^{\,i}}{i!} \tag{4}$$

$$p_t^a = p_t \cdot A_m \tag{5}$$

   where P_m is the probability of the tag sequence mutation based on the error rates, C_t is the tag count, S is the cumulative count of the tag cluster (including the parent tag count), and A_m is a Bonferroni-like *p*-value adjustment: $A_m = L \cdot 11$ for single-mutation tags and $A_m = 11 \cdot L \cdot \frac{11 \cdot L - 1}{2}$ for double-mutation tags (*see* **Note 6**).
   c. Statistical testing of individual tags. All tags with the adjusted *p*-value smaller than the *p*-value threshold are removed from the cluster (*see* **Note 7**).

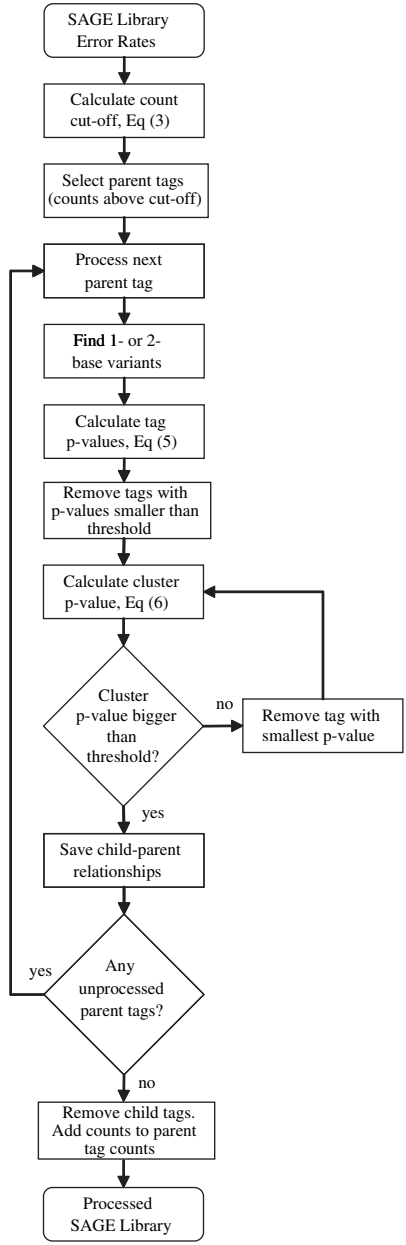


Fig. 3. Tag processing workflow.

d. The cluster *p*-value is calculated based on the counts of the mutant tags and the parent tag count (*see* **Note 5**):

$$p_c = \sum_{i=C_c}^{i=S} \frac{e^{-P_e} \cdot P_e^i}{i!} \tag{6}$$

where P_e is the probability of generating an erroneous tag sequence based on the error rates and C_e is the cumulative counts of the tags in the cluster excluding the parent tag count.

e. Statistical testing of the cluster size. If the cluster *p*-value is larger than the *p*-value threshold ($p_c > p_0$), then all the tags are retained in the cluster, otherwise the tag with the smallest *p*-value is removed from the cluster and the test is performed again.

f. Tag sequences retained in the cluster are called child tags and the child–parent relationships are saved for the final processing step.

3. When all the parent tags are processed, the child–parent relationships are resolved and the child tag counts are added to the parent tag counts with the child tag sequences deleted from the library (*see* **Note 8**).

4. Notes

1. In general, it is accepted that duplicate SAGE ditags are discarded. However, the expected proportion of ditags with two highly abundant tags may not be negligible. In this case, the tag counts may be adjusted after the ditag processing step.
2. It is important to optimize the search in the selection of one- or two-base variant sequences. For example, one might calculate the numbers of the nucleotides for ditag and tag sequences prior to the search and efficiently discard tag or ditag pairs when the numbers differ by three or more nucleotides.
3. **Eq. 1** is quickly solved with the secant method ***(6)***.
4. The error rate estimates in **eq. 2** were observed in a LongSAGE library of more than 100,000 tags. This library was preliminary processed by *phred* with the quality score of 20.
5. Poisson approximation to the binomial distribution is used in **eqs. 4** and **6**.
6. The Bonferroni-like adjustment constant in **eq. 5** represents the overall number of one-base mutant sequences and two-base mutant sequences. In the case of the one-base mutations, it is estimated as the number of bases (*L*) multiplied by the number of mutation types (11): 3 substitutions, 4 insertions, and 4 deletions. This number is certainly not precise but is a reasonable first order estimate.
7. The statistical tests in the tag processing step control for the type-II error (effectively the null hypothesis is if the observed variants are technology related mutants). As a result of the nature of this problem, exact calculation of the type-I *p*-value is not possible unless all genuine tag sequences are known.

8. When resolving child–parent relationships in the tag processing step, it is important to note that some of the tags may be classified both as child and parent, implying that the higher order relationships must be resolved.

References

1. Lash, A. E., Tolstoshev, C. M., Wagner, L., et al. (2000) SAGEmap: a public gene expression resource. *Genome Res*. **10,** 1051–1060.
2. Matsumura, H., Ito, A., Saitoh, H., et al. (2005) SuperSAGE. *Cell Microbiol*. **7,** 11–8.
3. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*. **20,** 508–12.
4. Akmaev, V. R. and Wang, C. J. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics* **20**, 1254–1263.
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**, 175–85.
6. Atkinson, K. E. (1988) *An Introduction to Numerical Analysis.* 2nd ed. John Wiley & Sons, New York, NY.

11

Duplicate Ditag Analysis in LongSAGE

Jeppe Emmersen

Summary

The Long serial analysis of gene expression (SAGE) protocol generates ditags from tags with overlapping overhangs, thereby increasing the probability of duplicate ditag formation in LongSAGE. In this chapter, a tool is presented that facilitates the analysis of duplicate ditags in LongSAGE studies to determine whether they should be included or not.

Key Words: SAGE; ditags; Perl; transcriptome; probability; PCR; amplification bias.

1. Introduction

In the original serial analysis of gene expression (SAGE) protocol, ditags were generated by ligating two blunt-ended 14-bp tags ***(1)***. To avoid bias due to cloning and PCR artifacts, duplicate ditags were discarded during the extraction process, a procedure that was continued after the LongSAGE protocol ***(2)***. By deriving the basic probability expressions for SAGE and LongSAGE, it is possible to verify whether discarding duplicate ditags has a significant effect on the resulting tag counts. Because the original SAGE protocol generates blunt-ended tags, the probability of two tags combining to form a ditag was independent of their sequence: $P(AB) = P(A)^*P(B)$, where the probability is simply the frequency of each tag. For two tags with equal probabilities of 0.02, the ditag probability would be 0.004. If 25,000 ditags were sampled to form a 50,000-tag library, the tag counts would be 1000 and the number of expected duplicate ditags would be 10—i.e., only 1 % of the tag counts. Discarding duplicate ditags, in this case, had little effect on the overall tag counts, and was thus justified by the risk that some

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press Inc., Totowa, NJ

may stem from experimental artifacts. In LongSAGE however, the *Mme*I enzyme generates a 21- to 22-bp long tag, which has a 2-bp overhang at the 3′ end. This overhang ensures that only tags complementary to each other at the 3′ end can ligate together. This changes the basic probability of each ditag AB to be chosen to: P(AB) = P(A)*P(B)*16, i.e., the probability of B is now dependent of A if A is chosen first and a uniform distribution of compatible overlaps is assumed. Going back to first example of a typical 50,000-tag SAGE study, two tags of tag count 1000 with a compatible overhang would, on average, give rise to 160 duplicate ditags, a nonnegligible 16 % fraction of the total tag count. In reality, the distribution of the 3′ overhangs is not uniform. **Figure 1** shows the distribution of overhang dinucleotides for a human pancreas SAGE library (Pa1b), which vary considerably from the uniform 1/16 distribution (=0.0625) ***(3)***. Using these data, up to a threefold difference in tag counts is observed, depending on inclusion or exclusion of duplicate ditags.

A plot of the relationship between tag counts generated by discarding or including duplicate ditag reveals no general nonrandom bias of including duplicate ditags (*see* **Fig. 2**). On the contrary, the inclusion of duplicates increases tag counts proportional with abundance, as should be expected from the probability equations.

A major complication of analysis is variability of the *MmeI* enzyme in the size of the restricted DNA fragments. The enzyme digests either 20 or 19 nucleotide downstream of its recognition sequence, generating two different overlaps, and consequently gives rise to 40, 41, or 42 bp ditags in different

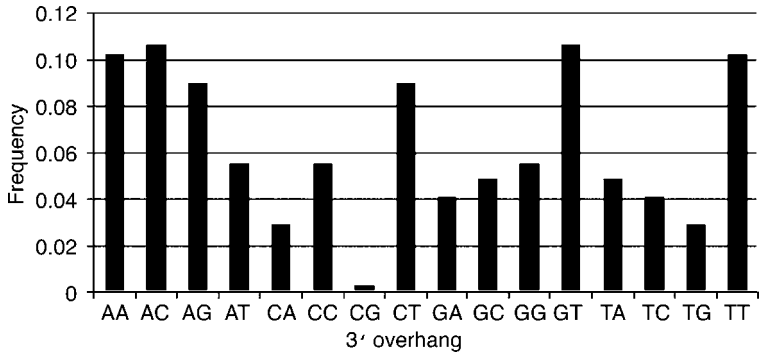


Fig. 1. Tags observed with and without inclusion of duplicate ditags from the LongSAGE study of pancreatic acinar cells ***(3)*** shows a linear relationship of tag counts.

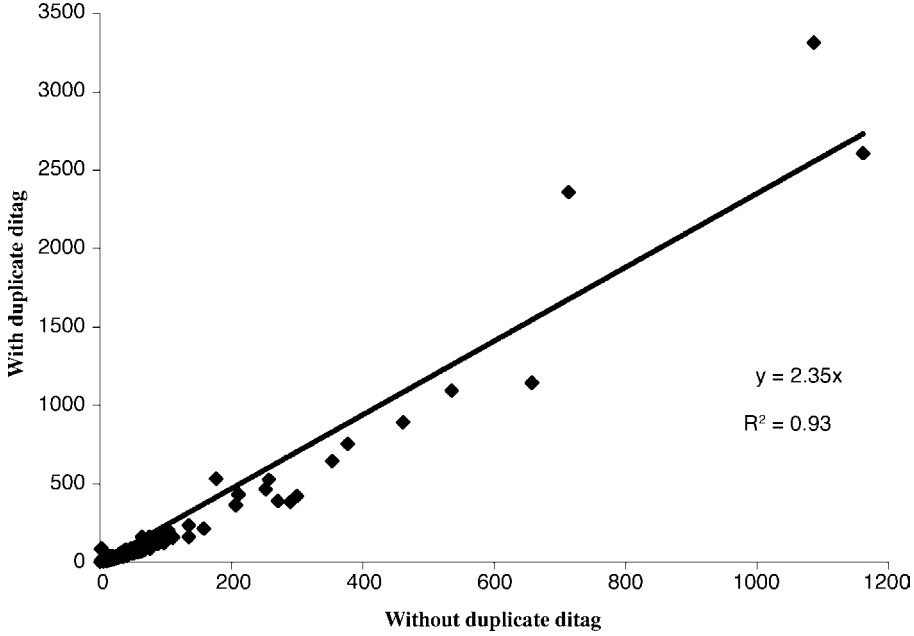


Fig. 2. The distribution of observed compatible overhangs from ditags of lengths 40 bp and 42 bp.

ratios (**Table 1**). Therefore, the ditag probability must be calculated for each individual fragment. Taking these considerations into account, the ditag probability becomes: $P(A'B') = T_{A'}/T_{total} * T_{B'}/T_{PPT}$, where $T_{A'}$ is total tag count of A′, $T_{B'}$ is total tag count of B′, T_{total} is total library tag count, and T_{PPT} is total possible partner tags for the overhang between A′ and B′, with A′ and B′ being one of two possible length representations of the tags. The expected occurrence of each ditag in library of D_{total} becomes: $D_{AB} = D_{total} * T_A/T_{total} * T_B/T_{PPT}$.

The last expression provides the theoretical basis to verify the observed duplicate ditag counts by correlating with the predicted ditag counts when

Table 1
The Distribution of Ditag Lengths in a LongSAGE Study of Pancreas Acinar Cells *(3)*

Tag length	35	36	37	38	39	40
Number of ditags	222	3342	11330	7307	218	53

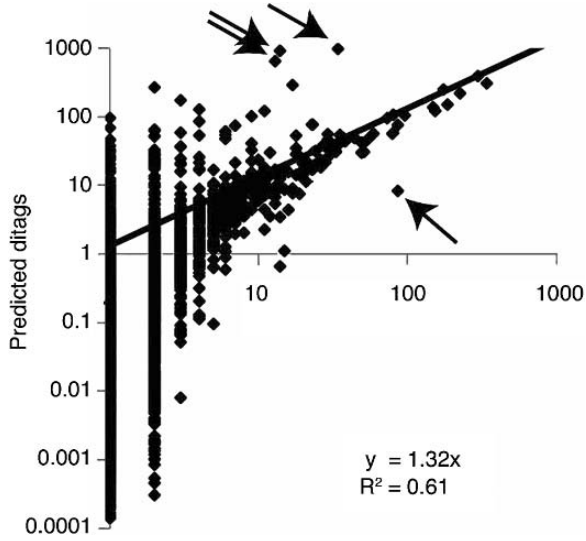


Fig. 3. Observed vs predicted ditags from the LongSAGE study of pancreatic acinar cells *(3)*. Outliers identified as contaminants are indicated by arrows. Removal of these increases the correlation coefficient from 0.61 to 0.95.

duplicate ditags are included in the tag extraction. Tags with large deviations from the predicted counts can be verified manually and possible contaminants removed from further analysis. For SAGE libraries generated from amplified mRNA, this may be a useful quality check of the amplification linearity *(3,4)*. This is implemented in the longsage_bias algorithm, which is implemented as a Perl program described in the methods section of this chapter. This algorithm can be used to correlate duplicate ditag counts and identify suspicious ditags that merit closer inspection, not to correct tag counts. An example is shown in **Fig. 3**, where the arrows indicate identified contaminants. Removal of these tags increases the Pearson Product Moment correlation from 0.61 to 0.95.

2. Materials

1. The analysis of duplicate ditags can be performed using the Perl script longsage_bias.pl. This script can be obtained from www.bio.aau.dk/en/biotechnology/software_applications or by email request to je@bio.aau.dk. The Perl package can be obtained from www.activestate.com for windows and is normally a standard on Unix machines. The module Getopt::Std is needed for the script to run and can be obtained from either ActiveState using the PPM engine or by searching CPAN.

Table 2
Output From the Predict File of the longsage_bias.pl Script

Tag	Overlap class	Count AB	A	B	P_long	Predicted
CATGTCAGGGTGATTCTGGTG	TGG	1306	3425	2467	0.934233	1762.88
CATGAATTGAAGAAACTGACC	CCT	1306	2467	3425	0.705714	1762.88
CATGGCGTGACCAGCTTTGTT	TTT	356	2669	1115	0.617424	327.1043
CATGGAACACAAAAAAAAAAA	AAA	356	1115	2669	0.382576	327.1043
CATGTTCATACACCTATCCCC	CCC	296	535	3425	0.710843	381.2448
CATGTCAGGGTGATTCTGGTG	TGG	296	3425	535	0.934233	381.2448
CATGTCCTCAAAACAAAAAAA	AAA	226	752	2669	0.409938	221.5748
CATGGCGTGACCAGCTTTGTT	TTT	226	2669	752	0.617424	221.5748
CATGGCGTGACCAGCTTTGTT	TTT	189	2669	534	0.617424	153.8708
CATGCTGAATCTAAATTATAA	AAA	189	534	2669	0.271028	153.8708
CATGAGCCTTGGTATCAAGAG	AGG	177	655	2467	0.532258	247.5677
CATGAATTGAAGAAACTGACC	CCT	177	2467	655	0.705714	247.5677
CATGTCCTCAAAAAAAAAAAA	AAA	157	434	2669	0.408602	127.8498
CATGGCGTGACCAGCTTTGTT	TTT	157	2669	434	0.617424	127.8498
CATGTCCTCAAAACAATAAAA	AAA	153	470	2669	0.495935	140.3755
CATGGCGTGACCAGCTTTGTT	TTT	153	2669	470	0.617424	140.3755
CATGTGCGAGACCACCCCTAT	ATT	105	898	1115	0.625	110.7055
CATGGAACACAAAAAAAAAAA	AAA	105	1115	898	0.382576	110.7055

2. The input files for this script is a directory of phd files generated by the Phred base caller ***(6)***.

3. Methods

To run the analysis, type **longsage_bias.pl** *Input_Directory*. This generates a new folder, containing all files generated by the analysis. The results folder is named for the input directory and the date of run.

The output from the analysis run consists of three files: a predict file, a ditag file, and a tag file. The analysis prints out several statistics—the log output. The log output can be redirected to a file using the < redirection operator.

1. The predict file is a tab separated file (*see* **Note 1**) listing all tags derived from duplicated ditags and various variables, the most important being the counts of two associated tag found in duplicated ditags and the predicted counts for each duplicated ditag (*see* **Note 2**) By plotting the observed ditag count against the predicted ditag count for each ditag pair, it is easy to spot any deviations from the linear relationship. In general, the less abundant a tag is, the more variation is seen as a consequence of sampling. As a rule of thumb, tag pairs exhibiting more than fivefold deviations from their predicted ditag count merit further investigation. For example, in a SAGE library from pancreas mRNA, an unknown ditag predicted to be found 8 times was found 86 times, a more than 10-fold increase in abundance. Further analysis by BLAST revealed this ditag to consist of two tags derived from the *Escherichia coli* β-lactamase gene, thus a likely result of contamination (*see* **Note 3**).
2. The log-file output contains several library-wide statistics from the ditag extraction and subsequent tag extraction **(Table 2)**. The first section lists possible monotags, i.e., where a single anchoring enzyme (CATG for *Nla*III) site was found in the beginning of the ditag sequence file, but no closing site within the chosen ditag size limit. The second section contains a resume of the ditag extraction: the number of total ditags found, the number of possible monotags, and the number of correctly formatted ditags rejected because of sequence quality. The third section is a list of ditags formed by the same monotag. The fourth section lists the total number of tags extracted from the ditag list, a distribution of the 16 possible dinucleotides participating in the overlap of the tags as generated by the *Mme*I enzyme (determined from ditags of length 40 and 42 bp), the number of total dinucleotide overlaps, and finally, a distribution of ditag lengths **(Table 1)**. The ditag file and the tag file are tabulated outputs of DNA tag sequences and corresponding tag counts. These may be used for further analysis, i.e., mapping of the tags.

4. Notes

1. Tabulator separated files are easily imported into any spreadsheet program, such as Excel, for further analysis.

2. For each ditag A-B, there are two entries: one for A-B and one for B-A. The values may differ slightly as a result of limitations in the data sets, but converge for larger tag counts.
3. To visualize as many outliers as possible, it is adventitious to plot the data using logarithmic axes.

References

1. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270(5235),** 484–487.
2. Saha, S., Sparks, A. B., Rago, C., et al. (2002) Using the transcriptome to annotate the genome. *Nat Biotechnol.* **20(5),** 508–512.
3. Heidenblut, A. M., Luttges, J., Buchholz, M., et al. (2004) aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Res.* **32(16),** e131.
4. Vilain, C., Libert, F., Venet, D., Costagliola, S., and Vassart, G. (2003) Small amplified RNA-SAGE: an alternative approach to study transcriptome from limiting amount of mRNA. *Nucleic Acids Res.* **31(6),** e24.
5. Dinel, S., Bolduc, C., Belleau, P., et al. (2005) Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res.* **33(3),** e26.
6. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998. **8(3),** 186–194.

12

Statistical Comparison of Two or More SAGE Libraries

One Tag at A Time

Gerben J. Schaaf, Fred van Ruissen, Antoine van Kampen, Marcel Kool, and Jan M. Ruijter

Summary

Several statistical tests have been introduced for the comparison of serial analysis of gene expression (SAGE) libraries to quantitatively analyze the differential expression of genes. As each SAGE library is only one measurement, the necessary information on biological variation or experimental precision is lacking. Therefore, each test includes its own approach to derive such a variance measure from the data set or a theoretical distribution. Because the confidence in tag counts depends on the library size, a test between two or more libraries should be based on original tag counts. When groups of libraries are compared, the test should determine that the proportion of a specific tag in all libraries is the same (null hypothesis), but also offer the possibility to detect specific differences between individual libraries and groups of libraries. The Z-test and the G-test encompass these characteristics and are described for the comparison of two libraries and (two or more) groups of libraries, respectively.

Key Words: SAGE-statistics; Z-test; binomial distribution; G-test; log–likelihood ratio; multinomial distribution; null hypothesis.

1. Introduction

The serial analysis of gene expression (SAGE) procedure can be described as taking a sample from a large population of tags and counting the number of each specific tag, and can thus be well approximated as sampling with replacement ***(1)***. The number of specific tags in a SAGE library per total number

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

of tags in that library is assumed to be an unbiased estimate of the number of copies of a specific mRNA per cell (*see* **eq. 1**). This fraction is, therefore, a direct estimate for the abundance or *proportion* of specific mRNA transcripts in that cell. The number of specific tags is the result of the probability of each tag to be sampled from the total population and, therefore, can be assumed to be binomially distributed. For the large number of tags sequenced in one library, this binomial distribution can be very well approximated by a normal distribution ***(2)***.

The aim of most SAGE studies is to identify genes of interest by comparing the number of specific tags found in two or more SAGE libraries. In other words, the aim of a comparison of SAGE libraries is to reject the null hypothesis that the observed tag counts in these libraries are equal. Testing of this hypothesis is hindered by the fact that each SAGE library is only one measurement: the necessary information on *biological* and *experimental* precision is lacking. Biological variation may reflect any biological influences, such as tissue-specific gene expression, changes in expression due to developmental differences, disease states, and transcripts lacking a particular recognition site for the anchoring enzyme (as a result of mutations or alternative splicing). Technical variation may be caused by incomplete digestion of samples leading to incorrect tags and tag counts, RNA degradation, or sequencing errors. Several statistical tests have been described as finding differences in tag abundance either between two (pair-wise) or among multiple libraries. These tests all include their own assumptions about the origin and properties of the tag variation. In general, the power of the statistical analysis of SAGE data depends largely on the accuracy of the determined tag proportion, which is dictated predominantly by the library size (for detailed description of size effects, *see* **ref. *2***). In this chapter, we describe the Z-test (comparison of two libraries ***[3]***) and the G-test (comparison of multiple libraries ***[4]***). These tests focus on proportions of specific tags in each library and thus account for (differences in) library size. It is important to note that in most comparisons between specific tags in SAGE libraries, there is no *a priori* knowledge of the direction of the effect. Therefore, all decision rules must be formulated to result in a two-sided test. Note that all available tests for the statistical comparison of two or more SAGE libraries must be executed for each individual tag. No tests for groups of tags or whole libraries have been developed.

The Z-test proposed by Kal et al. ***(3)*** is based on the normal approximation of the binomial distribution ***(3,5)***. The test statistic Z is calculated as the difference in proportions divided by the standard error of this difference when the null hypothesis is true **(eq. 2)**. The Z-statistic is approximately normally distributed

and can, therefore, be compared to the $\alpha/2$-percentile of the normal distribution. An additional advantage of the Z-test is that this method of testing provides a way to calculate the number of tags that must be sequenced to detect a difference as significant. Other tests to identify differences between two SAGE libraries have been published and are extensively reviewed and compared in recent literature ***(2**,**6**,**7)***. In general, all tests for the comparison of tag counts between two libraries lead to the same conclusion ***(2)***. Computer programs for the comparison of two libraries have been made available ***(2**,**6**,**7)***.

Because all SAGE libraries can be considered an unbiased representation of the transcriptome of the tissue that they are derived from, all available SAGE libraries can be compared directly. Each new library can be included in the statistical comparison of (a selection of) the many libraries that are now available in a number of public databases. Statistical comparison of such a set of libraries is not straightforward. Performing all possible pair-wise tests in a large collection of libraries is statistically invalid and would lead to unacceptable accumulation of false positives (Type I error). When a correction for multiple testing (*see* **Note 1**) is used, the resulting test is very conservative. A number of tests has been suggested recently for the comparison of groups of SAGE libraries. These include the *t*-test between the normalized tag counts in two libraries ***(8)*** or the Chi-square test among pooled tag counts ***(9)***. The shortcoming of the first approach is that such a t-test treats all tag counts as equally reliable, which ignores the fact that confidence depends on library size ***(2**,**10)***. On the other hand, pooling of tag counts completely ignores the between-library variation. Neither approach can, therefore, be considered valid for testing differences between groups of SAGE libraries. To overcome these shortcomings, a weighted t-test ***(10)*** and, more recently, logistic regression approaches, which can deal with more than two groups of SAGE libraries, were proposed ***(11**,**12)***. Each of these procedures deals with the within-library and between-library variation by introducing weight factors or scaling terms. In all of these methods, the testing procedure starts with the classification of the SAGE libraries into two or more groups based on *a priori* assumptions. However, similarly to other multiple-comparison-of-group approaches, an overall test of the null hypothesis that all libraries share the same tag proportion should precede such a grouping of libraries. To this end, we recently introduced a test that is the direct multilibrary extension of the Z-test ***(3)*** or Chi-square test ***(13)***. Like these two-sample tests, the multisample log–likelihood ratio test, (dubbed G-test in agreement with Sokal and Rohlf ***[14]***) is based on the binomial distribution of SAGE tags. The G-test, which originates from the analysis of frequencies, offers the above-mentioned overall test as well as a straightforward procedure by which

to continue when this null hypothesis is rejected, and it can be concluded that the tag proportions in the libraries are not all equal. The latter procedure was adapted to fit the design of SAGE experiments and has recently been described in detail ***(4)***.

2. Materials

2.1. Equations

2.1.1. Proportions

The proportion (p) of a certain mRNA transcript with tag count n in a SAGE library of size N:

$$p = \frac{n_{\text{specific tags}}}{N_{\text{total tags}}} \left(= \frac{n_{\text{specific mRNA/cell}}}{N_{\text{total mRNA/cell}}} \right) \tag{1}$$

The expected proportion in each of k libraries, when the null hypothesis, that all libraries have equal proportions, is true:

$$p_0 = \sum_k n_k \Big/ \sum_k N_k \tag{2}$$

2.1.2. Z-Test

Test statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2}} \tag{3}$$

with p_1 and p_2 from **eq. 1** and p_0 from **eq. 2**

Detectable difference:

$$p_1 - p_2 > Z_{\alpha/2}\sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2} + Z_{\beta}\sqrt{p_1(1-p_1)/N_1 + p_2(1-p_2)/N_2} \tag{4}$$

Required number of tags in each library:

$$N > \left(\frac{Z_{\alpha/2}\sqrt{2p_0(1-p_0)} + Z_{\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{p_1 - p_2} \right)^2 \tag{5}$$

2.1.3. G-Test

Overall test statistic:

$$G_{\text{intrinsic}} = 2\sum_{k}\sum_{s+ns}\{n \cdot \ln(n/n_0)\} \tag{6}$$

with n and n_0 per library.

Test statistic per individual library:

$$G_{\text{individual},t\to st} = 2\sum_{s+ns}\{n \cdot \ln(n/(p_{st}N))\} \tag{7}$$

Test statistic per group of libraries:

$$G_{\text{pooled},t\to st} = 2\sum_{s+ns}\left\{\sum_{k} n_k \cdot \ln\left(\sum_{k} n_k \Big/ \left(p_{st}\sum_{k} N_k\right)\right)\right\} \tag{8}$$

2.2. Software

The Z-test and G-test have been implemented in the SAGEstat and G-test software programs, respectively. SAGEstat accepts manual input of data, but for complete libraries, both programs are based on data stored in Microsoft Excel. The programs are available on request by sending an e-mail to biolab-services@amc.uva.nl with subject SAGEstat or Gtest, respectively.

3. Methods

In this section, we describe how to perform the calculations required for the Z-test and the G-test, as they are implemented in the above-mentioned programs. The Z-test is to be used to test differences between tag counts in two libraries **(Subheading 3.1)**, and can be used to plan a SAGE experiment with two libraries **(Subheading 3.2)**. For the comparison of tag counts in groups of multiple libraries, the log–likelihood ratio or G-test is to be used **(Subheading 3.3)**. Note that these test procedures between libraries are performed for each tag individually.

3.1. Comparison of Tag Counts in Two Libraries: Z-Test

3.1.1. Difference between Two Tag Counts

1. Calculate the proportion (p_1 and p_2) of the tag in each library **(eq. 1)**.
2. Calculate the proportion p_0, which is the expected proportion when the null hypothesis is true **(eq. 2)**.

3. Use these proportions, and the library sizes N_1 and N_2 in **eq. 3** to calculate the test statistic Z.
4. Compare Z to the $\alpha/2$ percentile of the normal distribution ($Z_{\alpha/2}$) or calculate the P-value of Z using the NORMSDIST function of MS Excel.

In the SAGEstat program, the library sizes and specific tag counts can be entered manually and the program will perform these calculations **(Fig. 1)**.

For the comparison of all tag counts in two libraries, the previous procedure is to be repeated for every tag. This procedure has been automated for all tags in two libraries and implemented in the SAGEstat software. Note, however, that by testing tags in the libraries sequentially, the Type I error accumulates (*see* **Note 1**). A multiple testing correction must be applied to prevent false positives.

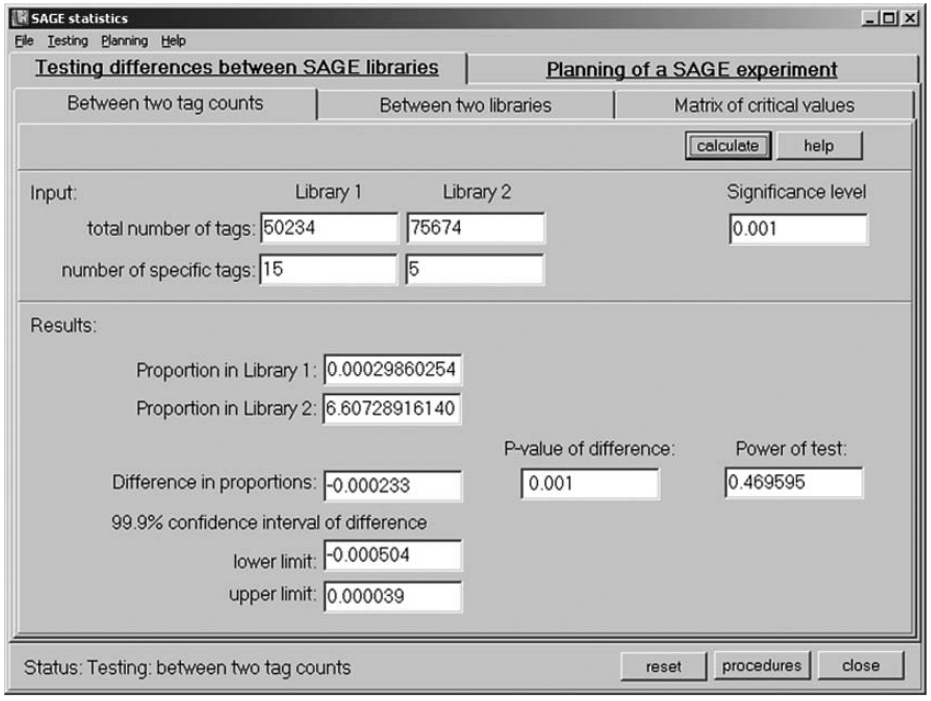


Fig. 1. The master screen of SAGEstat. SAGEstat can be used to perform the different possibilities of the Z-test (testing and planning experiments with two SAGE libraries) as described under **Subheadings 3.1.** and **3.2.**

3.1.2. Calculation of Critical Values

Another way to easily compare the tag counts in two whole libraries is to calculate critical values, which are defined as the tag counts that must be found in the second library for the difference with the tag count in the first library to be statistically significant different. The critical values depend on the library sizes N_1 and N_2, and the calculation requires the user to determine the accepted level for the Type I error (α, incorrect rejection of the null hypothesis; two-sided) (*see* **Note 2**).

The SAGEstat program can be used to perform these calculations. In principle, the program systematically simulates increasing tag counts in the first library and uses **eq. 3** (with increasing tag counts) to determine the number of tags in the second library at which the resulting *P*-value leads to rejection of the null hypothesis at the required level of significance.

3.2. Planning a SAGE Experiment With Two Libraries

The normal approximation of the binomial distribution that forms the basis of the Z-test can also be used to easily calculate the required number of tags in each library for a certain expected difference in proportion to be tested as significantly different. The principle of these calculations is given in **Fig. 2**. The

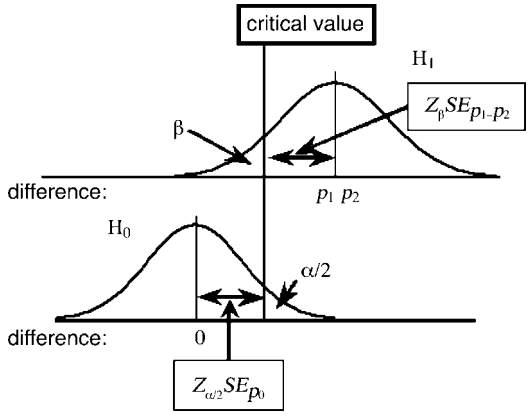


Fig. 2. Illustration of the derivation of the sample size equation. When the null hypothesis (H_0) is true, the expected value of the difference $p_1 - p_2$ is 0. The distance between this expected value and critical value is $Z_{\alpha/2}SE_{p0}$. When the alternative hypothesis is true, the distance between the true $p_1 - p_2$ and the critical value is $Z_\beta SE_{p1-p2}$. Therefore, the distance from 0 to $p_1 - p_2$ is given by the sum of these two distances **(eq. 4)**.

calculation can be performed manually as described under **Subheading 3.2.1.**, but for a range of differences they can easily be carried out using the SAGEstat software. The calculations are based on **eq. 4**, which gives the difference in proportions that can be expected to be detected at significance level α and with a power 1−β for libraries of size N_1 and N_2. For equal library sizes, **eq. 4** can be converted into **eq. 5** ***(2,5)***. In this form, the equation can be used to calculate the common library size that is required to detect an expected difference ($p_1 - p_2$) at significance level α with a power 1−β.

3.2.1. Calculate the Required Number of Tags in Each Library

1. Determine the expected difference in tag proportion.
2. Decide which significance level (α) is required and look up $Z_{\alpha/2}$ in the normal distribution.
3. Decide which power (1−β) is required and look up Z_{β} in the normal distribution.
4. Use **eq. 5** to calculate the library size *N* for both libraries.

The SAGEstat program can be used to calculate the library size for a range of expected differences, defined as fold-difference between tag proportions **(Fig. 3)**.

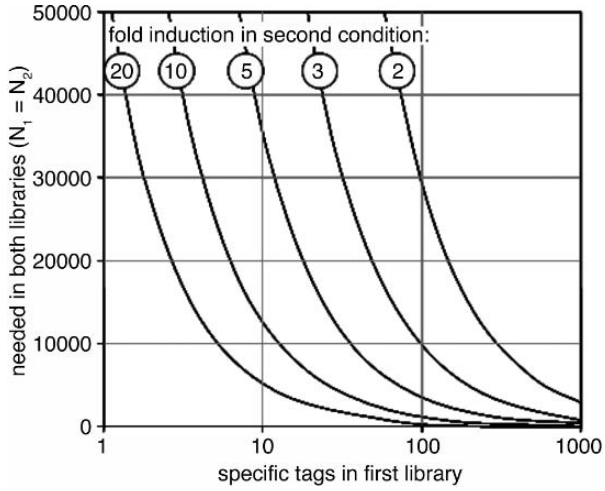


Fig. 3. Planning (SAGE) experiments: number of tags that must be sequenced in two SAGE libraries. Number of tags that must be sequenced in each of the libraries to detect a 2- to 20-fold difference in abundance at a significance level of 0.001 and a power of 0.9.

While planning a SAGE experiment, **eq. 4** can also be used to:

- Calculate the detectable difference between two libraries when the library sizes, the significance level, and the required power have been determined.
- Calculate the reachable power of the experiment for an expected difference when the library sizes and the significance level have been decided.
- Calculate the number of tags required in the second library when the number in the first library, the significance level, and the required power are fixed.

The latter option can be used when one is planning to compare a new library to a SAGE library that has already been constructed. Because the conversion of **eq. 4** into **eq. 5** is not possible when the library sizes N_1 and N_2 are not equal, these calculations must be based on **eq. 4.** This requires an iterative procedure that is much too laborious to execute manually, and a program such as SAGEstat must be used. Examples of such calculations are given in Ruijter et al., 2002 ***(2)***.

3.3. Comparison of Tag Counts in Multiple Libraries: G-Test

As discussed in the Introduction, the analysis of multiple SAGE libraries comprises a two-step procedure. First, the null hypothesis that the proportion for a specific tag is the same in all libraries in the set must be tested. To this end, the overall *G*-statistic ($G_{intrinsic}$; **eq. 6**) is calculated and compared to the Chi-square distribution. When this null hypothesis is rejected, different approaches can be chosen to determine in which libraries the tag is differentially expressed. One can calculate an individual *G*-statistic for each library to test the significance of its deviation from the overall expected proportion. However, after the rejection of the null hypothesis, the results of such a test are hard to interpret biologically. When *a priori* knowledge of the included libraries is available, e.g., based on tissue origin, one can choose to continue with a comparison of the user-defined standard and test subsets of libraries. This approach is referred to as an *"a priori"* or "supervised" comparison of subsets. Together with the overall test, the proposed test procedure proceeds through five decision rules, and a tag must pass each subsequent rule to be considered differentially expressed between the standard and test set of libraries.

The calculation of the *G*-statistic is based on a multinomial distribution, which is a generalization of the binomial distribution to the case where an attribute has more than two classes. For the calculation of the *G*-statistic, the data for one tag are placed in a contingency table with the libraries as columns and two rows per column. The top row of this table contains the tag counts for the current tag in each library (specific counts) and the bottom

row contains the number of other tags (nonspecific counts.) The numerator of the likelihood ratio is the probability of observing the frequencies in this contingency table. The denominator is this probability when the libraries and tag counts are independent. The *G*-statistic is defined as two times the logarithm of this likelihood ratio and, for a comparison of k SAGE libraries, simplifies to **eq. 6** ***(14)***.

3.3.1. Rule 1: Overall Test

First, test the overall null hypothesis that all k libraries in a collection of libraries have the same proportion of a specific tag. Because the overall *G*-statistic is based on the expected proportion, which is calculated from all data in the current data set, this statistic is referred to as $G_{intrinsic}$ **(Fig. 4)**.

1. Calculate p_0, which is the expected proportion when the null hypothesis is true, using **eq. 2**.
2. Use p_0 and each library size to calculate n_0; the number of expected tags in that library when the null hypothesis is true is $n_0 = p_0 N$, in which N is the size of the library.
3. Calculate the "$n \cdot \mathrm{Ln}(n/n_0)$" term in **eq. 6** (in the case of zero tag counts, *see* **Note 3**). In this term, *Ln* represents the natural logarithm, and n is the number of observed tags in a library. This term must be calculated and summed for specific and nonspecific tags. This sum is the contribution of each library to $G_{intrinsic}$.
4. Calculate $G_{intrinsic}$ according to **eq. 6** (sum over all k libraries and multiply by 2)

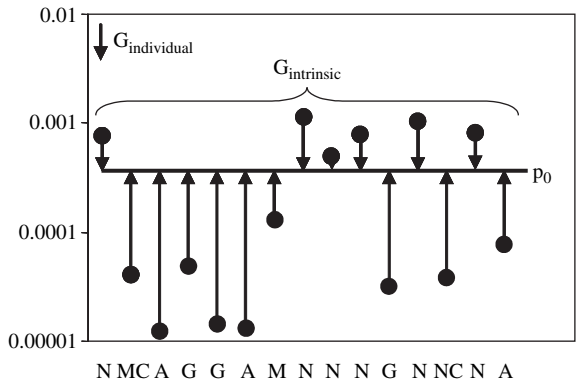


Fig. 4. Testing the null hypothesis with the G-test (overall test; $G_{intrinsic}$). $G_{intrinsic}$ is used to test the homogeneity of the whole library set (rule 1). This test uses the intrinsic p_0 as reference proportion. Note that $G_{intrinsic}$ is the sum of $G_{individual}$.

The distribution of the *G*-statistic approximates the Chi-square distribution ***(14)*** and, therefore, the *P*-value of the observed $G_{intrinsic}$ is determined from the χ^2-distribution with $k-1$ degrees of freedom (df). Based on the chosen significance level α (*see* **Note 2**) and the obtained *P*-value, the null hypothesis is rejected when $G_{intrinsic} > \chi^2_{\alpha,n-1,df}$. In that case, the tag can be considered to be informative, i.e., the information contained in the tag counts may reflect some aspects of the difference between the original libraries **(Fig. 4)**.

When $G_{intrinsic}$ leads to rejection of the overall null hypothesis, one can proceed to the next step **(Subheading 3.3.2.)** of the G-test procedure. To this end, the libraries are grouped into a standard set and one or more test sets (*see* **Note 5** for guidelines).

3.3.2. Rule 2: Homogeneity of the Standard Set

The intrinsic *G*-statistic of the standard set is calculated ($G_{within,\ st}$) to test the homogeneity among tag proportions found in the standard set.

1. Calculate the expected proportion for the standard set (p_{st}; **eq. 2** with the sums of *n* and *N* of the libraries in the standard subset; **Figs. 5** and **6**).
2. Calculate $n_{0,st}$ for each library in the standard set.
3. Calculate $G_{within,\ st}$ similarly to $G_{intrinsic}$ (**eq. 6** with *k* as the number of libraries in the standard set).

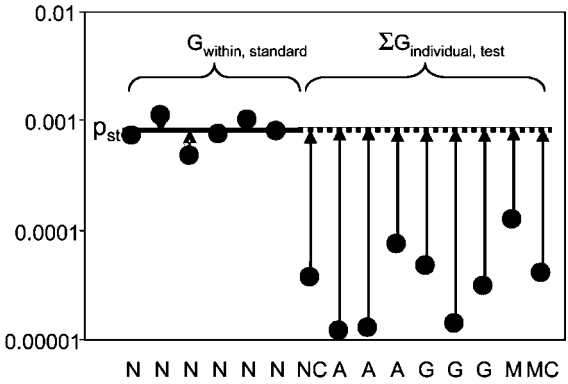


Fig. 5. Supervised G-test: $G_{within,\ st}$ and $G_{individual,\ t->st}$. Supervised G-test with the proportion of an user-defined standard subset (p_{st}) as reference. $G_{within,\ st}$ is used to test the homogeneity of the standard set of libraries (rule 2). $G_{individual,\ t->st}$ tests the deviation of each individual library in the test set from the average proportion of the standard set (rule 3).

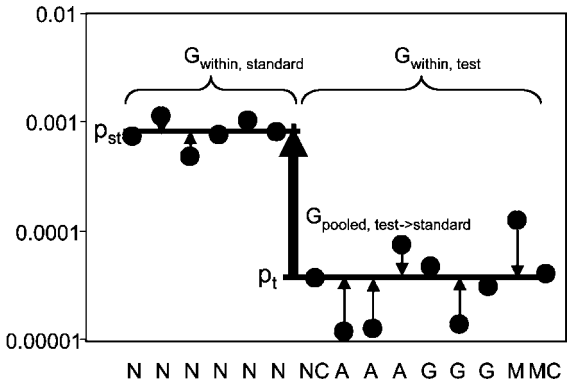


Fig. 6. Supervised G-test: $G_{within,\ st}$, $G_{within,\ t}$ and $G_{pooled,\ t->st}$. Supervised G-test with the proportions of an user-defined test subset (p_t) and standard subset (p_{st}) as reference. $G_{pooled,\ t->st}$ is used to test the difference between a test set and the standard set (rule 4). $G_{within,\ t}$, tests the homogeneity of the test set and can be used to obtain a more stringent set of differentially expressed genes (rule 5). The sum of $G_{individual,\ t->st}$ adds up to the sum of $G_{pooled,\ t->st}$ and $G_{within,\ t}$

Like $G_{intrinsic}$, the degrees of freedom of $G_{within,\ st}$ are the number of libraries in the subset minus 1. When the tag proportion in the standard set is considered homogeneous ($G_{within,\ st} < \chi^2_{\alpha,n-1,df}$), proceed to next rule **(Subheading 3.3.3).** Otherwise, the tag is considered to be not informative for the current subset classification.

3.3.3. Rule 3: Compare Individual Libraries in the Test Sets With the Standard Set

To determine whether the tag proportion in each individual library in the test set can be considered significantly different from the overall abundance of the standard set, the $G_{individual,t->st}$ is calculated for each test library.

1. Calculate the proportion p_{st} of the standard subset (as under **Subheading 3.3.2.**)
2. Calculate the expected tag count in the each test library, $n_{0,st}$, as the product of p_{st} and the size N_t of the current test library **(Fig. 5)**
3. Calculate $G_{individual,\ t->st}$ using **eq. 7**. $G_{individual,\ t->st}$ is two times the sum of $n_t \cdot \ln(n_t/n_{0,\ st})$, with n_t as the number of observed tags in the current test library. This term is summed for the specific and nonspecific $(s + ns)$ tags.

Each $G_{individual,\ t->st}$ has one degree of freedom. When the tag proportion in each individual library in a test set is considered significantly different from the overall abundance of the standard set ($G_{individual,\ t->st} > \chi^2_{\alpha,1,df}$) one can proceed to next rule **(Subheading 3.3.4).** Otherwise, one has to consider this tag noninformative. This

rule may be too stringent when many libraries are included in the comparison. In that case, it may be necessary to change the stringency of the G-test (*see* **Note 6**).

3.3.4. Rule 4: Test the Difference Between a Test Set and the Standard Set

To determine whether the average tag proportion of a test set libraries differs significantly from that of the standard set the $G_{pooled,\ t->st}$ is calculated. $G_{pooled,\ t->st}$ reflects the difference between the average proportion observed in the test set and the expected proportion when the test set does not differ from the standard set **(Fig. 6)**.

1. Calculate the *pooled* observed tag count n_t for the subset with k libraries.
2. Calculate the pooled expected tag count $n_{0,st}$ using the expected proportion p_{st} and the sum of the library sizes N_t.
3. Calculate $G_{pooled,\ t->st}$ using **eq. 8**.

When more than one test set is defined, $G_{pooled,\ t->st}$ must be calculated for each set. $G_{pooled,\ t->st}$ has one degree of freedom. When $G_{pooled,\ t->st} > \chi^2_{\alpha,1,df}$, the tag is considered to have a significant different proportion in the test set compared to the standard set. Optionally, to increase the stringency for selecting tags with significant different proportions, proceed to next rule **(Subheading 3.3.5).**

3.3.5. Rule 5: Homogeneity of a Test Set of Libraries

To obtain a more stringent set of tags with significantly different proportions, one can calculate the $G_{within,\ t}$. This G-statistic reflects the homogeneity of proportions of a specific tag in the test subset of n libraries **(Fig. 6)**. Tags that pass this rule can be considered to have homogeneous abundances within the test set of libraries. $G_{within,t}$ is calculated similarly to $G_{intrinsic}$ (**eq. 6** with k as the number of libraries in the test subset) but with n_0 based on the expected proportion of the test subset (**eq. 2** with sums of test subset). Like $G_{intrinsic}$, the degrees of freedom of G_{within} are the number of libraries in the subset minus 1. When the G-statistic is significant ($G_{within,\ t} < |\chi^2_{\alpha-1df}$), the hypothesis that the test set is homogeneous must be rejected.

For the relation between the different G-statistics, consult **Note 4**. This supervised G-test procedure was described in detail in Schaaf et al. ***(4)***, and examples of tags that are either rejected by or pass each of the five rules are shown in **Table 1**. The above G-test procedure is implemented into the G-test computer program **(Fig. 7)** ***(4)***.

Table 1
Application of the Decision Rules on 25 Tags From Four Rhabdomyosarcoma (RMS) and Two Normal Muscle SAGE Libraries *(4)*

					ARMS36 RMS	ERMS112 RMS	ERMS12 RMS	ERMS102 RMS	Musle_old Normal	muscle_yng Normal		
library > type >>			Gintrinsic	Gwithin standard	Gindividual standard						Gpooled t->st	Gwithin test
FinalMC (UC 163)	Gene symbol	TAG										
MC		AAAAATAAAG	0.000	0.000	0.001	0.237	0.028	0.764	0.000	0.003	0.111	0.001
Hs.415722	LOC283120	AAAGAAATGG	0.000	0.047	0.000	0.160	0.590	0.840	0.258	0.102	0.000	0.000
MC		AACCAAAAAA	0.014	0.008	0.013	0.934	0.442	0.426	0.036	0.102	0.585	0.070
Hs.436439	PTK9L	AACCTGGCCT	0.044	0.023	0.069	0.031	0.302	0.548	0.069	0.168	0.003	0.953
Hs.183435	NDUFB1	AAGAATCTGA	0.001	0.033	0.000	0.041	0.024	0.439	0.109	0.162	0.000	0.248
MC		AAGACAGTGG	0.000	0.009	0.000	0.405	0.551	0.154	0.058	0.069	0.000	0.000
		AAAAAAAAAA	0.000	0.994	0.000	0.000	0.000	0.000	0.996	0.996	0.000	0.000
		AAAAAACATT	0.048	0.994	0.824	0.916	0.000	0.498	0.996	0.996	0.002	0.044
Hs.432491	ESD	AAAAAACTCC	0.001	0.974	0.069	0.002	0.000	0.186	0.982	0.982	0.000	0.002
		AAAACAGTGG	0.017	0.202	0.000	0.409	0.338	0.906	0.307	0.446	0.004	0.033
		AAAACATTCT	0.000	0.051	0.000	0.000	0.000	0.000	0.165	0.171	0.000	0.424
Hs.387804	PABPC1	AAAAGAAACT	0.000	0.994	0.000	0.000	0.000	0.000	0.996	0.996	0.000	0.000
		AAAATACTGA	0.048	0.994	0.824	0.001	0.061	0.001	0.996	0.996	0.000	0.181
		AAAATGAAAA	0.007	0.994	0.824	0.916	0.000	0.498	0.996	0.996	0.000	0.009
		AAAATGTACT	0.008	0.994	0.009	0.000	0.002	0.498	0.996	0.996	0.000	0.446
Hs.63657	NGLY1	AAAATTATCT	0.026	0.994	0.824	0.000	0.613	0.001	0.996	0.996	0.000	0.104
Hs.133892	TPM1	AAAGTCATTG	0.000	0.201	0.000	0.000	0.000	0.000	0.363	0.370	0.000	0.886
Hs.255950	LOC154866	AAATGTGCTG	0.000	0.682	0.000	0.000	0.002	0.030	0.776	0.769	0.000	0.953
Hs.446354	TCEA3	AACAAGGTGA	0.000	0.942	0.000	0.000	0.041	0.017	0.959	0.959	0.000	0.512
MC		AACCCAGGAG	0.000	0.329	0.000	0.000	0.000	0.050	0.518	0.465	0.000	0.753
MC		AACCCGGGAG	0.000	0.320	0.000	0.000	0.000	0.000	0.467	0.497	0.000	0.488
		AAGCTGAGGT	0.000	0.485	0.000	0.000	0.000	0.000	0.664	0.585	0.000	0.004
Hs.375921	RPL31	AAGGAGATGG	0.000	0.804	0.000	0.000	0.000	0.000	0.862	0.859	0.000	0.000
Hs.405590	EIF3S6	AATATTGAGA	0.000	0.672	0.000	0.000	0.000	0.000	0.773	0.756	0.000	0.027
Hs.433394	TUBA3	AATGCTTTGT	0.000	0.994	0.009	0.000	0.000	0.000	0.996	0.996	0.000	0.000
		Rule >		1	2		3				4	5

The cells in the table give the *P*-values of the respective G-statistics. The decision rules are indicated at the bottom of the table. Grayed-out *P*-values do not satisfy the decision rules. Only the tags without gray cells are considered to be differentially expressed. All rules are tested with $\alpha = 0.05$ as significance level.

4. Notes

1. The Type I error accumulates when more than one hypothesis test is done within one experiment. When n pair-wise comparisons are done and each decision is based on a significance level α, then the chance that one or more of the decisions are wrong accumulates to $1-(1-\alpha)^n$, which for 10 comparisons and $\alpha = 0.05$, already increases to over 40%. Therefore, the significance level (α) should be adjusted to safeguard against the accumulation of false-positive rate that may result from multiple testing. Several approaches to correct for multiple testing have been described, such as the Bonferroni correction ***(5)*** or the False Discovery Rate (FDR; ***15***). With the thousands of tests in one SAGE analysis, the Bonferroni correction becomes very stringent and therefore very conservative. Other methods have been proposed to remedy this. Van den Oord and Sullivan ***(16)*** argue that it is better to base the corrected significance level on reducing false discoveries as

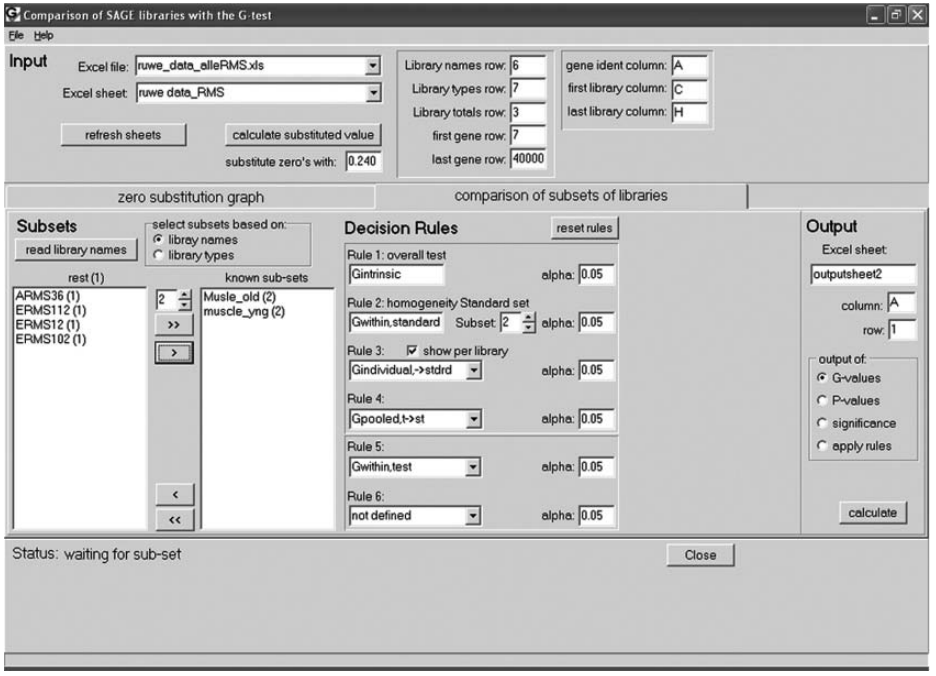


Fig. 7. G-test program interface to define the five decision rules as described under **Subheading 3.3.**

well as the proportion of true detection and give an equation to this end ***(16)***. A common misconception is that the Type I error occurs when the test is performed. When one has 10 tags with the same tags counts in two libraries, of course only one test is be performed and the results copied for each of the 10 tags. However, when it comes to accumulation of Type I error one has in fact done all 10 tests. This is because the error does not occur during the test but during the sampling: each of the 10 samples may contain a random sampling error and, therefore, turn out to be a false positive.

2. The significance level α of a hypothesis test is the chance that one is willing to accept that a true null hypothesis is rejected (Type I error or false positive). In *discovery-driven* research, the choice of α is not immediately important. The tags that are the most promising for future research are the ones that deviate most from the expected proportion, and these are the ones with the lowest *P*-value. So, ranking the tags in ascending order of *P*-value provides a list with the most interesting genes at the top. A very different situation occurs when research is *hypothesis-driven* or diagnostic. Then, the choice of α determines whether or not a tag is labeled "differentially expressed." In short, the choice of α depends on

the consequences of a false-positive decision on the one hand and those of a false-negative decision on the other.

3. In the statistical analysis of SAGE libraries, tags that are not observed in one of the libraries included in the test are often excluded from the test. However, very low abundant transcripts, and thus zero tag counts, still may reflect a true low transcript abundance, and their exclusion from the analysis may constitute a loss of valuable information. However, the $n \cdot \ln(n/n_0)$ term that is part of the equations of all G-statistics can not be calculated when the observed tag count n is zero. Therefore, a zero-substitution procedure must be implemented when comparing multiple libraries. Replacing zeros by 1 might lead to high tag abundances in small libraries compared to the tag count of 1 in a larger library. Similarly, replacing zero tag counts by a number close to zero can lead to clearly false-positive test results. The zero-substitution procedure that is implemented must have a minimal effect on the test statistic and should not contribute to the decision whether to reject the null hypothesis. To obtain such an optimal zero substitution value, one has to simulate random tag counts of 0 and 1 for each of the libraries in the current collection of SAGE libraries. These simulated tag counts can be generated in such a way that the tag count is always 0 in the smallest library and always 1 in the largest library. The other libraries are assigned either a 0 or 1, with the chance of getting a 1 depending on the library size. After each assignment of a 0 or a 1 tag count to each of the libraries, an iterative procedure determines a zero-substitution value that gives the minimal $G_{intrinsic}$ for this simulated tag. The whole process must be repeated and the mean zero-substitution value is then used in the G-test procedure of all tags in the libraries. This way, it is ensured that the zero-substitution, a value between 0 and 1, in itself will add only a minimal contribution to the G-statistic. This approach is purposely conservative: genes that change from no to low expression and in which the tag count difference between 0 and 1 is real will undeservedly not show up as differentially expressed. However, the described substitution procedure itself will not lead to a significant *P*-value and therefore avoids false positives.
4. An important characteristic of the *G*-statistic is the additivity of G-values. The total G-value in the supervised comparison of subsets is given by the following equation.

$$\sum G_{individual} = \sum G_{within} + \sum G_{pooled} \quad (9)$$

This additivity of *G* is also observed at the subset level: the sum of its $G_{individual}$ values is equal to the sum of G_{within} and the G_{pooled} of that set. The G_{pooled} of the standard set is 0. Note that the above sum of $G_{individual}$, also denoted as $G_{extrinsic}$, is greater than $G_{intrinsic}$ **(eq. 6)**. The latter is based on the overall proportion of all libraries, whereas $G_{extrinsic}$ is based on the average proportion of the standard subset. Because $G_{extrinsic}$ is based on an *a priori* defined set, its degrees of freedon are equal to the number of libraries.

5. The supervised procedure as described under **Subheading 3.3.** can be summarized as a comparison of the mean tag proportion in the test set with the mean proportion in the standard set. Therefore, the most reliable selection of the differentially expressed tags is obtained when the tag proportions in the standard set are highly homogeneous, and the mean tag proportion in the standard set is a good representation of the abundance of the transcript in that group of libraries. In other words, the variation in the standard set needs to be minimal. When comparing SAGE libraries from diseased tissue to libraries generated from their normal counterparts, generally, the variation between normal tissues from different individuals will be smaller than the variation between the diseased tissues from these individuals. These normal tissue SAGE libraries should then be grouped in the standard set, those from the diseased tissue in the test subset. However, when comparing, for instance, subtypes of tumors, this distinction will not be so clear. One can approach this problem by performing a hierarchical cluster analysis (*see,* for instance, Shannon et al. ***[17]***) using different distance measures. Then, the standard set can be chosen as those libraries that cluster most closely together. It should be noted that hierarchical cluster methods are developed for continuous data and are not specifically developed for the discrete data that SAGE data represents. Cluster analysis will be highly influenced by low tag counts, and such tags should not be included when this analysis is performed. Another drawback of cluster analysis of SAGE data is the data normalization that must precede such a test. Because the confidence in SAGE tag counts depends on the library size, normalized data can only be compared when the library sizes are similar.
6. The supervised procedure as described under **Subheading 3.3.** was optimized for a comparison of six libraries (two in standard set and four in the test subset). These subsets of libraries were very different so that the G-test readily identified 277-tags as significantly different at a significance level of 0.05. However, when the number of libraries increases, rule 3 ($G_{\text{individual, } t->st} > \chi^2_{\alpha,1,\text{df}}$) may be too stringent. One can lower the stringency of this step in the G-test procedure by setting a threshold on the number of libraries in the test set that must pass rule 3 **(Subheading 3.3.3.)**. As mentioned above, rule 5 **(Subheading 3.3.5.)** is optional and confers much stringency to the G-test. Obviously, the number of tags with significantly different proportions selected is dependent on the chosen significance level for each of the rules. Note in this respect that rules 1, 2, and 5 test homogeneity whereas rules 3 and 4 test a difference (*see also* **Note 2**).

References

1. Stollberg, J., Urschitz, J., Urban, Z., and Boyd, C. D. (2000) A quantitative evaluation of SAGE. *Genome Res.* **10,** 1241–1248.
2. Ruijter, J. M., Van Kampen, A. H., and Baas, F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics* **11,** 37–44.

3. Kal, A. J., van Zonneveld, A. J., Benes, V., et al. (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell* **10,** 1859–1872.
4. Schaaf, G. J., Ruijter, J. M., van Ruissen, F., et al. (2005) Full transcriptome analysis of rhabdomyosarcoma, normal, and fetal skeletal muscle: statistical comparison of multiple SAGE libraries. *FASEB J.* **19,** 404–406.
5. Altman, D. (1991) *Practical Statistics for Medical Research*. Chapman-Hall, London: pp. 161–167, 253–258.
6. Man, M. Z., Wang, X., and Wang, Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* **16,** 953–959.
7. Audic, S. and Claverie, J. M. (1997) The significance of digital gene expression profiles. *Genome Res.* **7,** 986–995.
8. Ryu, B., Jones, J., Blades, N. J., et al. (2002) Relationships and differentially expressed genes among pancreatic cancers examined by large-scale serial analysis of gene expression. *Cancer Res.* **62,** 819–826.
9. Walter-Yohrling, J., Cao, X., Callahan, M., et al. (2003) Identification of genes expressed in malignant cells that promote invasion. *Cancer Res.* **63,** 8939–8947.
10. Baggerly, K. A., Deng, L., Morris, J. S., and Aldaz, C. M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19,** 1477–1483.
11. Baggerly, K. A., Deng, L., Morris, J. S., and Aldaz, C. M. (2004) Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics* **5,** 144.
12. Lu, J., Tomfohr, J. K., and Kepler, T. B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6,** 165.
13. Michiels, E. M., Oussoren, E., Van Groenigen, M., et al. (1999) Genes differentially expressed in medulloblastoma and fetal brain. *Physiol. Genomics* **1,** 83–91.
14. Sokal, R. R. and Rohlf, F. J. (1995) Analysis of frequencies, in *Biometry, the Principles and Practice of Statistics in Biological Research,* 3rd edition. W. H. Freeman and Co.: New York: pp. 685–789.
15. Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.* **9,** 811–818.
16. van den Oord, E. J. and Sullivan, P. F. (2003) False discoveries and models for gene discovery. *Trends Genet.* **19,** 537–542.
17. Shannon, W., Culverhouse, R., and Duncan, J. (2003) Analyzing microarray data using cluster analysis. *Pharmacogenomics* **4,** 41–52.

13

Scaling of Gene Expression Data Allowing the Comparison of Different Gene Expression Platforms

Fred van Ruissen, Gerben J. Schaaf, Marcel Kool, Frank Baas, and Jan M. Ruijter

Summary

Serial analysis of gene expression (SAGE) and microarrays have found a widespread application, but much ambiguity exists regarding the amalgamation of the data resulting from these technologies. Cross-platform utilization of gene expression data from the SAGE and microarray technology could reduce the need for duplicate experiments and facilitate a more extensive exchange of data within the research community. This requires a measure for the correspondence of the results from different gene expression platforms. To date, a number of cross-platform evaluations (including a few studies using SAGE and Affymetrix GeneChips) have been conducted showing a variable, but overall low, concordance using different overall correlation approaches, such as Up/Down classification, contingency tables, and correlation coefficients. Here, we demonstrate an approach to compare two platforms based on the calculation of the difference between expression ratios observed in each platform for each individual transcript. This approach results in a concordance measure per gene, as opposed to the commonly used overall concordance measures between platforms. This between-ratio difference is a filtering-independent measure for between-platform concordance. Moreover, the between-ratio difference per gene can be used to identify transcripts with similar regulation on both platforms.

Key Words: Serial analysis of gene expression; SAGE; microarray; GeneChips; gene expression; expression profiling.

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

1. Introduction

Methods for the analysis of gene expression profiles have gone through progressive development over recent years. Traditionally, the level of transcribed mRNA has been analyzed using methods such as Northern blots, quantitative reverse-transcription (RT)-PCR, differential display ***(1,2)***, representational difference analysis ***(3)***, total gene expression analysis ***(4)***, and suppressive subtractive hybridization ***(5,6)***. All of these methods, although fruitful and still in use, have a limited scope with regard to the number of genes that can be analyzed simultaneously. Because of this limitation, new methods have been developed, including serial analysis of gene expression (SAGE) ***(7)***, massive parallel signature sequencing (MPSS) ***(8)***, cDNA and oligo microarray chip technologies ***(9–13)***, and Affymetrix GeneChips ***(11)***.

SAGE is based on the high-throughput sequencing of concatemers of short (13–14 bp; recently, 21–25 bp) sequence tags that originate from a known position within a transcript and therefore theoretically contain sufficient information to identify a transcript ***(7)***. In contrast to microarrays, SAGE estimates the abundances (expression levels) of thousands of transcripts without prior knowledge of the transcripts being expressed. The proportion of the tag within the total number of tags in the library gives a direct estimate of the abundance of the transcript within a biological sample. The proportional nature of the data enables easy exchange among researchers, thus allowing the creation of large public SAGE data sets for numerous human tissues, both normal and diseased ***(14,15)***.

In contrast to SAGE, DNA microarrays are used to measure relative expression levels between samples of thousands of known transcripts. Currently, three array variants are being used (for reviews, *see* **refs. *16*** and ***17***), i.e., spotted cDNA microarrays, spotted oligonucleotide microarrays, and synthesized oligonucleotide microarrays (Affymetrix GeneChips). The advantages of Affymetrix GeneChips are that they are highly sensitive, enabling the detection of mRNAs present at levels as low as 1 transcript in 100,000 ***(11)*** when the probe-labeling step is not considered ***(18)***. They are suitable for high-throughput analyses of multiple samples, and data can easily be shared and used for comparisons by other researchers using the same chips.

At present, SAGE, oligo microarrays, complementary DNA (cDNA) microarrays, and Affymetrix GeneChips are the most widely used techniques for determining gene expression levels and gene expression ratios. These different gene expression profiling platforms are often used in parallel, and data generated with the different techniques must be compared, and interchanged, within and between laboratories.

To determine the overall correspondence between expression levels or expression ratios of two different platforms, several methods have been used **(Fig. 1A–C)**. These include the parametric (Pearson) or nonparametric (Spearman) correlation coefficients between platforms, and contingency tables with varying numbers of classes for each platform. For the latter, a correspondence measure can be calculated as the percentage of transcripts falling in the cells on the diagonal **(Fig. 1B)**. An extreme form of the contingency table has only two classes per platform (ratios above and ratios below 1) and therefore only four cells. This form of concordance estimation is dubbed "Up/Down classification" **(Fig. 1A)**. None of these correspondence measures was deemed satisfactory because they treat very different ratios as similar (points A and B in **Fig. 1A**), which disqualifies the Up/Down classification as a reliable agreement measure. The use of contingency tables with more classes is a better approach, but here, some genes will be considered to be "in disagreement" although they have nearly corresponding expression ratios in both platforms (points A and B in **Fig. 1B**). The Pearson correlation coefficient is a measure for the fraction of variation in Y that is explained by the variation in X, and as such, only gives a measure for the tendency of the plotted points to increase simultaneously (solid line, **Fig. 1C**). However, when studying the correspondence between gene expression platforms, the expected linear relation has a slope of 1, when the results of both platforms are in complete correspondence (dashed line, **Fig. 1C**), and the deviation of the observed scatter plot from this expected relation should be tested. To remedy these pitfalls, we recently introduced a correspondence measure based on the difference between the log(ratio) values in the two platforms for each individual transcript ***(19)*** **(Fig. 1D)**. Apart from serving as the basis for a measure for overall platform concordance, this method also provides the user with an agreement measure for each individual transcript, which is of more interest than the overall correlation.

To demonstrate our approach, between-ratio difference measures were compared to the customary correlation measures in a comparison of the SAGE and Affymetrix platforms using reference RNA and Wilms' tumor gene expression ratios ***(19)***. The comparison of gene expression ratios based on contingency tables with Up/Down classification and a contingency table diagonal lead to an agreement of 63 % and 76 % between platforms, respectively. The Pearson correlation coefficient between platforms was 0.453 ($P < 0.01$). Regression analysis shows a linear trend with a slope of 0.477 for Affymetrix v SAGE, which, according to the correlation coefficient, differs significantly from a slope of 0. However, this slope also deviates significantly from the slope value of 1, which is expected when the platforms are identical (t-test for slopes; $P < 0.001$). A comparison of SAGE and Affymetrix data

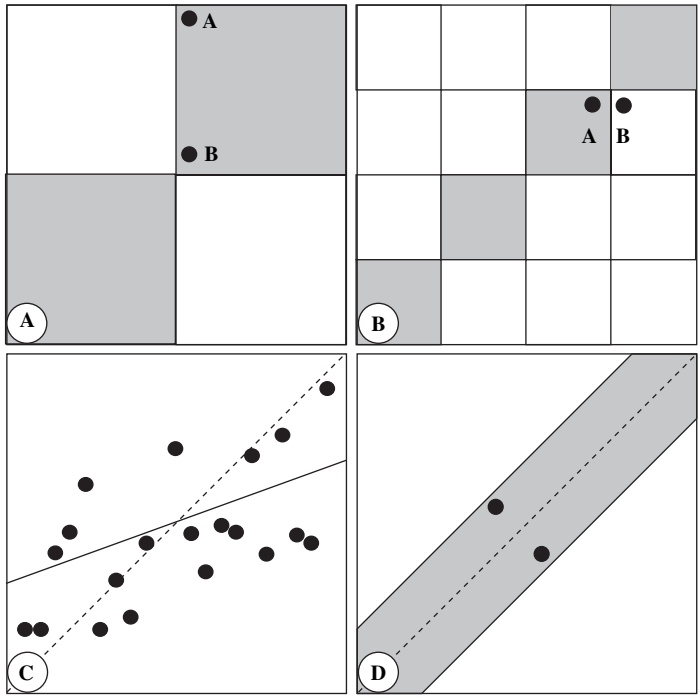


Fig. 1. Illustration of the methods used for the comparison of expression profiles from different platforms. **(A)** Up/Down classification: the points A and B, with very different ratios, are both considered to reflect a common tendency. **(B)** Contingency table diagonal: the points A and B, with very similar ratios, end up in different classes. **(C)** Correlation coefficients: the solid line fits to the point cloud which has a significant correlation coefficient between X and Y. However, the dashed line ($Y = X$) is the expected line when both platforms show identical expression patterns. **(D)** Schematic representation of the proposed method where the agreement measure between platforms is based on the absolute difference between expression ratios.

using our proposed classification based on the difference between the two ratios per Unigene cluster and accepting a 0- to 3-fold difference as indicative for agreement between the two platforms (red points in **Fig. 5**, discussed later) showed that the two platforms have an agreement of 78 %. This correspondence measure turned out to be hardly sensitive to different selections of Unigene clusters. Splitting the data based on the expression level resulted in 78 % and 90 % agreement for low and high expressed genes, respectively. This shows that the between-ratio difference results in a robust between-platform correspondence measure. Moreover, the between-ratio difference provides the user

with a correspondence measure per individual gene that can be used to select those genes for which a predetermined correspondence level is reached.

The overall similarity, in our example, between SAGE and Affymetrix GeneChips is modest when expression ratios are compared. The correspondence improves to 90 % when only highly expressed transcripts are included, which means that noise is filtered out for both platforms. The differences between SAGE and Affymetrix GeneChips were not caused by a biased selection of the final data set, differences in GC content of the included transcripts, or extreme ratios resulting from low gene expression values. The observed cross-platform differences arise from intrinsic properties of the platforms themselves, differences in the principle of determining the expression levels, such as absolute (SAGE) vs quantitative (microarray) mRNA levels, and/or processing and analytical evaluation ***(20)***.

In this chapter, we describe the application of our approach to determine the similarity between SAGE- and Affymetrix GeneChips-generated gene expression profiles of two independent RNA samples.

2. Materials

To perform the comparison of SAGE and Affymetrix (or microarray) platforms using the approach outlined in the methods section, one should either obtain SAGE and Affymetrix data of a test (X) sample and a control (Y) sample (e.g., National Center for Biotechnology Information [NCBI] website; Gene Expression Omnibus) or create the data using the I-SAGE kit (Invitrogen) and standard manufacturer's protocols for Affymetrix or microarray hybridizations. Ultimately, one will end up with two data sets consisting of an identifier in the case of SAGE a tag; for Affymetrix, a probe ID or accession number; and the respective gene expression data of a test and a control sample. In this chapter, we describe the comparison of SAGE and Affymetrix platforms (which has been described in full detail in **ref. *19***). Through the Methods section and supplementary notes, we give results from our comparison to demonstrate our findings. For a comparison of two other platforms, different results will be obtained.

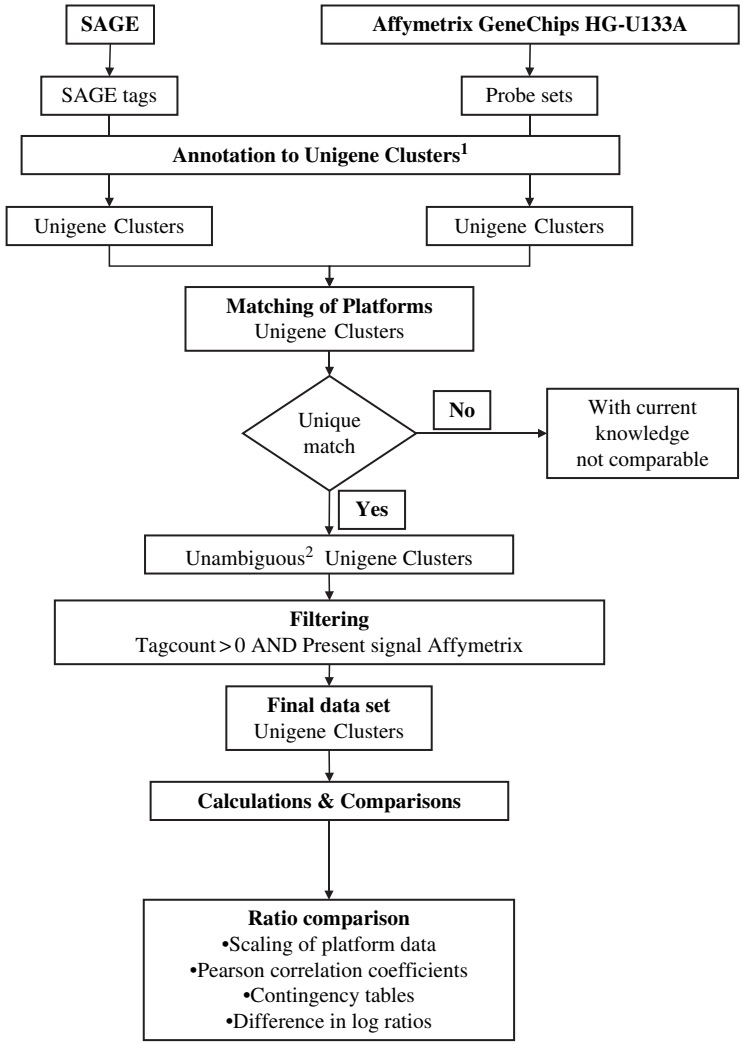
3. Methods

3.1. Annotation of SAGE Tags and Affymetrix Probe Sets

1. Annotate the extracted SAGE tags based on the SAGE Genie principles ***(21)*** (*see also* Chapter 15) through several stringent filters using data from the Cancer Genome Anatomy Project (CGAP) website ***(15)***. A detailed explanation is given in **ref. *19***. In the final comparison, tags matching to multiple Unigene clusters must be excluded.
2. Match the probe sets on the Affymetrix chips to their corresponding Unigene clusters based on the accession numbers provided by Affymetrix.

3.2. Matching of Platforms

1. Match the data from the two different gene expression profiling platforms on the basis of their Unigene Cluster (*see* **Fig. 2** and **Note 1**).
2. Include only those clusters for which a one-to-one relation between the two platforms is found. These clusters are called unambiguous Unigene clusters (*see* **Note 2**).
3. Filter the data for the presence of gene expression (tag count > 0 in both SAGE libraries and a "present" signal on the arrays for both RNA samples) (*see* **Fig. 3** and Note **3**).
4. Calculate the between sample ratio for each gene in each platform = X/Y (*see* **Note 4**).



5. Check that the final data set is an unbiased representation of the total dataset, and verify that the dataset is not skewed by ratios obtained through low expression signals or GC content (*see* **Note 5**).

3.3. Scaling of Expression Ratios Between Platforms (see Fig. 4)

The relation between the gene expression ratios observed in the SAGE and Affymetrix platforms cannot be assumed to be a simple, linear, $Y = X$ relation. This will already be clear from the different ranges of ratio values in each platform. To compare the ratios observed in both platforms, at least the range of observed ratios should be similar. In this scaling, the 10th and 90th percentiles are used to prevent undue influence of extreme ratios. Because in each platform, the observed ratio of 1 can be assumed to be true (*see* **Note 6**), the simplest function to scale the range of ratios of one platform to that of the other platform is a quadratic equation based on three values.

1. Calculate the ^{10}Log of the between-sample ratios for each gene for each platform.
2. Calculate the 10th percentile of the log(ratio)s less than 0 ($= R_{10}$) and the 90th percentile of the log(ratio)s above 0 ($= R_{90}$) for each platform.
3. Calculate the parameters for a quadratic scaling of platform B (Affymetrix) to the 10th–90th percentile range of platform A (*see* **Note 6**).

$$\begin{aligned} a &= 0 \\ b &= (R_{90,A}.{R_{90,B}}^2 - R_{10,A}.{R_{90,B}}^2)/(R_{90,B}.{R_{10,B}}^2 - R_{10,B}.{R_{90,B}}^2) \\ c &= (R_{90,A}.R_{10,B} - R_{10,A}.R_{90,B})/({R_{90,B}}^2.R_{10,B} - R_{90,B}.{R_{10,B}}^2) \end{aligned} \quad (1)$$

Fig. 2. Flow chart for matching data from two gene expression platforms. Serial analysis of gene expression (SAGE) tags were converted into Unigene clusters using data from the Cancer Genome Anatomy Project website. Accession numbers from Affymetrix GeneChips were also converted to their corresponding Unigene cluster. Platforms are matched according to their Unigene cluster, and only unambiguous Unigene clusters are selected. Finally, data are filtered for tag counts >0 and present calls on microarray platforms.

[1]In the complete process of annotation, a large number of tags or probe sets are lost for the following reasons. SAGE: 11,733 tags with no annotation, 13,113 tags with no reliable annotation, 913 tags with multiple Unigene Clusters, 80 tags belonging to linker sequences, 20 tags belonging to repetitive sequences, 22 tags belonging to mitochondrial DNA. Affymetrix: 1795 Probe sets no longer belong to a Unigene Cluster (Build 160). The remaining 20,488 probe sets represent 13727 unique Unigene clusters.

[2]Unambiguous Unigene clusters refer to those clusters that occur only once within each platform.

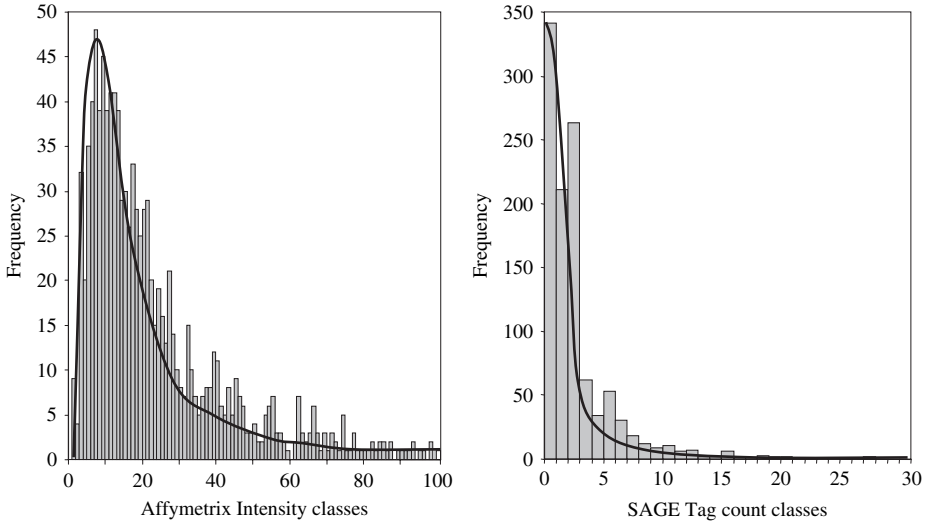


Fig. 3. Demonstration of the unbiased sample selection. Frequency distribution of the Affymetrix intensity and serial analysis of gene expression (SAGE) tag counts from the final matched data set (1094 Unigene clusters) and the total matching data set. The smoothed line represents the distributions of the total data set in each platform. For both Affymetrix (classes with an intensity width of 10) and SAGE (classes based on tag counts), the distributions of the final data set and the total data set do not differ from each other (Chi-square values of 327 [df = 323; $P = 0.412$] and 104 [df = 105; $P = 0.506$], respectively).

4. Use parameters a, b, and c to calculate the scaled log(ratio) of all log(ratio) values of platform B (Affymetrix):

$$R_{B,Scaled} = a + b.R_B + c.R_B^2. \quad (2)$$

3.4. Overall Comparison of Expression Ratios Between Platforms

After transforming both platforms to the same scale, the absolute difference between the log(ratio)s per individual gene can be calculated. The resulting between-ratio differences can be classified into classes of width 0.5, which corresponds to an approximate threefold difference in expression ratio between platforms (*see* **Note 7**). The resulting classification can be used to calculate a correspondence measure between platforms. Additionally, these classes can be used to label individual genes in a between platform scatter plot of ratios.

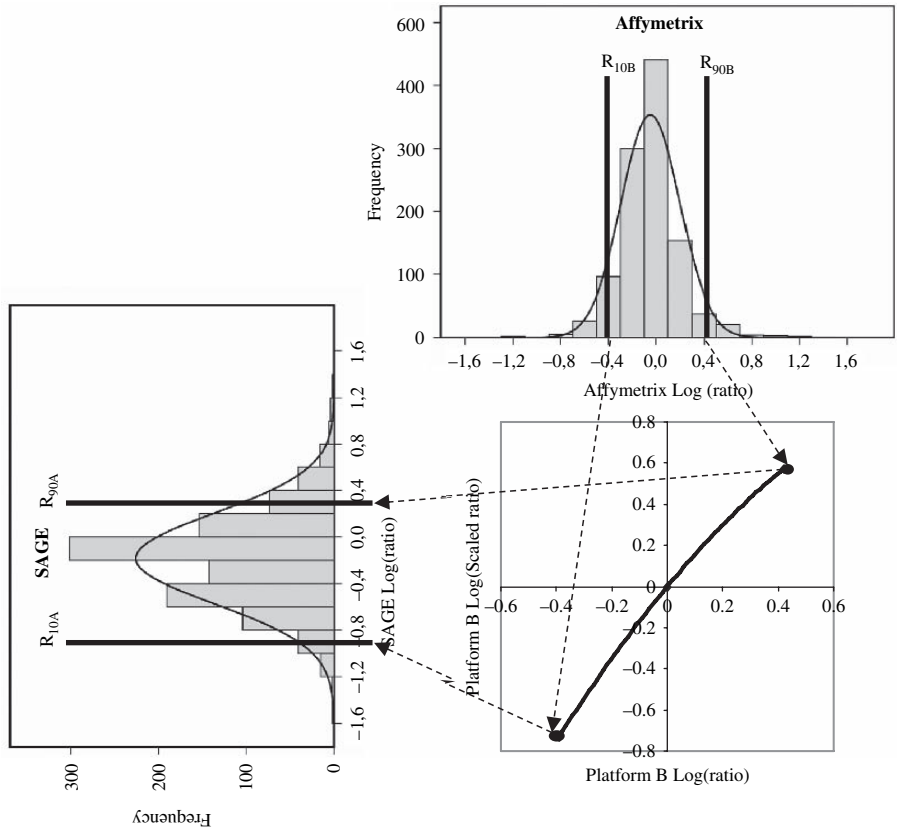


Fig. 4. Example of the scaling procedure. This figure represents the steps necessary to scale one platform to the other. Here, we scaled the Affymetrix platform to the serial analysis of gene expression (SAGE) platform as outlined in this chapter. On top and on the left, the Log ratios are drawn for both platforms. In the center, the graph of the quadratic equation is plotted, and the dashed arrows represent the scaling of the Affymetrix platform to the SAGE platform using the calculated 10th and 90th percentiles.

1. Calculate the between-ratio difference for each gene:

$$\text{Difference}_i = \text{ABS}(\text{R}_{\text{A},i} - \text{R}_{\text{B,scaled},i}) \tag{3}$$

2. Recode the between-ratio differences into classes [ClassNDiff_i; (0 through 0.5 = 0.5) (0.5 through 1 = 1) (1 through 1.5 = 1.5) (1.5 through 2 = 2) (2 through 2.5 = 2.5) (2.5 through 3 = 3) (3 through 3.5 = 3.5) (3.5 through Highest = 4)] (*see* **Note 7**).

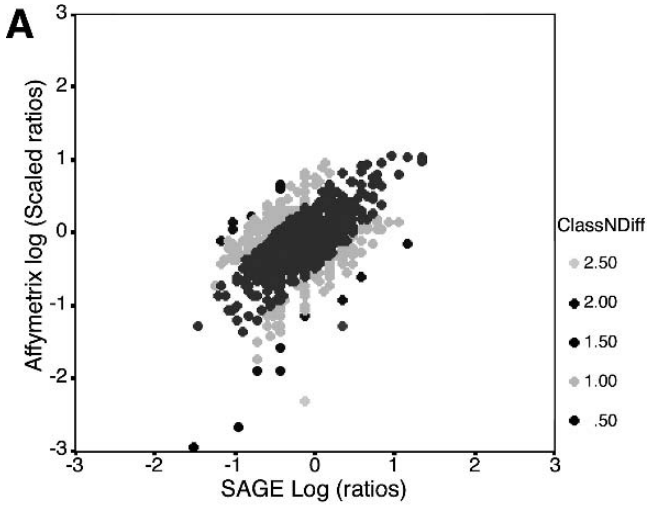
3. Create a frequency table of the resulting classes and calculate the correspondence measure between platforms as:

$$\text{Correspondence} = 100^{*}\text{class}(0.5)/\text{number of ratios} \quad (4)$$

4. Create a scatter plot of the log(ratio) values of platform A against the scaled log(ratio)values of platform B and mark the spots by their class (*see* **Fig. 5**)

3.5. Comparison of Expression Ratios of Individual Genes Between Platforms

Because the between-ratio differences are approximately normally distributed **(Fig. 6)**, it is possible to attach a probability value to each individual



B ClassNDiff

Classes	Frequency	Percent
0 – 0.5	852	77.9
0.5 – 1.0	225	20.6
1.0 – 1.5	14	1.3
1.5 – 2.0	2	.2
2.0 – 2.5	1	.1
Total	1094	100.0

Fig. 5. Concordance measures between platforms. **A** and **B** represent an example of how to visualize the concordance measure of the two platforms. **A** shows a graph in which the different concordance classes are plotted in different colors, and **B** shows a frequency table of the different concordance classes.

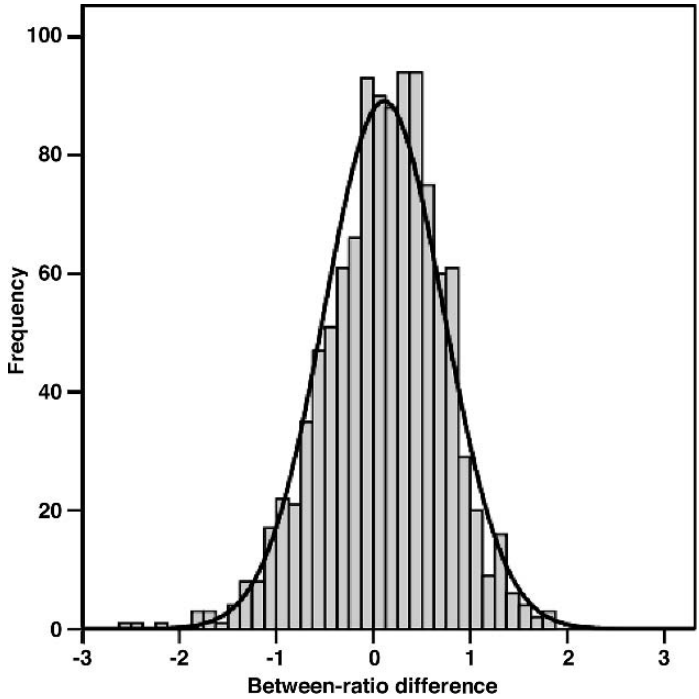


Fig. 6. Concordance measure per gene. The frequency distribution of the between-ratio differences is shown, and from the plotted normal curve, it can be seen that the between-ratio differences have an approximately normal distribution. This can, as explained in detail, be used to calculate a probability value for each individual between-ratio difference.

difference. This *P*-value gives an indication of the position of a gene in the between-ratio difference distribution and as such can be used to identify genes with a deviating behavior between platforms.

1. Calculate a standardized difference ($Diff_{St}$) for each gene *i*. $Diff_{St,i} = ((R_{A,i} - R_{S,i}) - m) / s$; where m is the mean of the between-ratio differences and s is the standard deviation of the between-ratio differences (*see* **Note 8**).
2. Use a normal distribution function to assign a *P*-value to this standardized difference.

Note that this *P*-value cannot be used to test whether the ratio difference equals zero. Such a test requires a gene-specific variance estimate in the denominator of the standardized difference, and such a variance estimate cannot be obtained from the four nonreplicated expression values that are used to calculate the ratio difference (*see* **Note 9**).

4. Notes

1. The matching of data from two different gene expression profiling platforms (as illustrated in **Fig. 2**) poses a few problems. In our example, on the one hand, a SAGE tag may link to more than one Unigene cluster, which results in matches with multiple different Affymetrix probe sets. On the other hand, multiple tags originating from one Unigene cluster might match with one Affymetrix probe set. Examining all multiple matches for each individual transcript is extremely laborious and beyond the scope of this study.
2. This matching step already results in a considerable reduction of data available for the comparison.
3. Our final example comparison of SAGE and Affymetrix contains 13 % of the SAGE Unigene clusters and 8 % of the Affymetrix Unigene clusters. These data represent 32 % of the unambiguous Unigene clusters. Only 1094 tags and probe sets were uniquely matched to the same Unigene clusters and were "present" in both tissue samples and platforms. This relatively low number underscores the major problem of how to merge different expression platforms. However, in view of the following quantitative comparison of gene expression platforms, it is important to note that a comparison of frequency distributions of all clusters and of the selected clusters showed that the final selection of 1094 Unigene clusters does not represent a biased sample, neither for the SAGE tag counts nor for the Affymetrix array intensities. This is illustrated in **Fig. 3**, in which the frequency distributions are given for Affymetrix intensities and SAGE tag counts from the final data set of 1094 Unigene clusters. The smoothed line, which represents the frequency distribution of all SAGE tag counts and all Affymetrix intensity data (only present calls), does not differ from the distribution of the subset included in the comparison of the two platforms.
4. In most gene expression studies, alterations of expression levels are expressed in relation to the simultaneously determined expression level of a reference sample, and conclusions are drawn based on these ratios. To this end, expression ratios were calculated between the reference RNA and the Wilms' tumor data for the SAGE tag counts as well as for Affymetrix HG-U133A GeneChips spot intensities. In this comparison, the final data set containing only the between-sample ratios for unambiguous transcripts was used, allowing effective comparison of the two platforms. This use of ratios might have the disadvantage of losing information about individual expression values. However, it corrects for platform specific variations, such as probe design, hybridization efficiencies, etc.
5. Variation due to "noisy fold ratios" generated from low-intensity transcripts is a widespread cause of error when computing statistics on ratios without accounting for the intensities from which the ratios were derived ***(22)***. Within our data set, we have shown that the final data set is an unbiased selection of the total data set **(Fig. 3)**. Additionally, the mean intensity signals for both SAGE and Affymetrix GeneChips appear to be randomly distributed over the ratio distribution (data not

shown). This indicates that the difference in expression ratios between platforms is not caused by low-intensity values. It has been suggested that the GC content of the transcripts could influence the correspondence between platforms ***(23)***. Statistical analysis showed that ratio differences did not depend on the GC content of the transcript. However, Unigene clusters showing good agreement between platforms tend to depend on the high GC content of the corresponding probe sets. This indicates that expression data from probe sets with a higher GC content show a better agreement with their corresponding SAGE data and are more reliable. Note in this respect that for a Unigene cluster, the GC content of a probe set is not necessarily the same as that of a transcript.

6. As a result of the chemistry, physics, and statistics of the detection technique, in each platform the observed gene expression is a nonlinear transformation of the real gene expression level. As an example, the saturation of the array hybridization means that the high expression levels are truncated. However, because such artifacts affect genes in both tissues in the same way, an observed expression ratio of 1 can still be expected for genes that are not differentially expressed in the studied tissues. On the other hand, these saturation effects, as well as the relatively larger Poisson error in the detection of low-intensity values, will affect the ratios on both sides of the ratio distribution in an unpredictable way. Similarly, the sampling error in SAGE will affect ratios for lowly expressed genes, despite the fact that SAGE tag counts are linearly related to transcript abundance. Finally, the discrete nature of tag counts, combined with the necessary normalization of tag counts to tags per 50,000, will have nonlinear effects on the observed ratio distribution in the SAGE platform. Therefore, the relation between the gene expression ratios observed in the SAGE and Affymetrix platform cannot be assumed to be a simple linear relation. To directly compare the ratios observed in both platforms, at least the range of observed ratios should be similar. Because in each platform the observed ratio of 1 can be assumed to be true, the simplest function to scale the range of ratio of one platform to that of the other platform is a quadratic equation based on the ratio of 1 and the 10th and 90th percentiles (R_{10} and R_{90}, respectively). The implementation of this quadratic scaling takes into account that the ratio distribution is not symmetrical around ratio 1. Note that the scaling uses log(ratio) values. Effectively, the ratio data of the Affymetric platform are scaled in such a way that: $R_{10B,scaled} = a + b.R_{10A} + c.R_{10A}^2$ and $R_{90B,scaled} = a + b.R_{90A} + c.R_{90A}^2$. Because ratio 1 stays 1, the intercept parameter (a) of the quadratic equation is 0. The above scaling rules result in the parameter equations given in **Subheading 3.3., step 4**. A schematic representation of the scaling is given in **Fig. 4**. Some of the choices in the scaling procedure can be considered to be *ad hoc*. However, given the current state of understanding of the causes for within and between platform variability, it was deemed best to opt for a simple quadratic scaling equation to convert the distribution of ratios, which is asymmetric around 1 to a common scale. With increasing knowledge on the physics, chemistry, and sampling statistics, better conversion functions will emerge.

7. The category width was arbitrarily defined at 0.5, representing a 0- to 3-fold difference. This width can be defined by the user to accommodate personal preference.
8. The approximately normal distribution of the between-ratio differences **(Fig. 6)** allows the calculation of a standardized difference value for each gene from which a *P*-value can be obtained. The standardized difference and its *P*-value can be used as a measure for the position of a specific gene within the distribution of between-platform ratio differences, and as such they can serve as a statistical threshold to determine which genes can be confidently interchanged between platforms. For instance, in the current study, the transcripts with a less than 0.5-fold between-ratio difference (red dots in **Fig. 5**) have a chance of at least 0.8 that they show similar gene expression on both platforms. Note, however, that this *P*-value cannot be used to test whether the ratio difference equals zero. Such a test requires a gene-specific variance estimate in the denominator of the standardized difference, and such a variance estimate cannot be obtained from the four nonreplicated expression values that are used to calculate the ratio difference.
9. For the comparison of two different gene expression profiles starting from ratio measurement for both platforms, a software application is available to perform all calculations en produce an output of calculated values (E-mail: biolab-services@amc.uva.nl; Subject, PlatformScaling).

References

1. Liang, P. and Pardee, A. B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971.
2. Martin, K. J. and Pardee, A. B. (1999) Principles of differential display. *Methods Enzymol*. **303**, 234–258.
3. Lisitsyn, N. and Wigler, M. (1993) Cloning the differences between two complex genomes. *Science* **259**, 946–951.
4. Sutcliffe, J. G., Foye, P. E., Erlander, M. G., et al. (2000) TOGA: an automated parsing technology for analyzing expression of nearly all genes. *Proc. Natl. Acad. Sci. USA* **97**, 1976–1981.
5. 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.
6. Wang, X. and Feuerstein, G. Z. (2000) Suppression subtractive hybridisation: application in the discovery of novel pharmacological targets. *Pharmacogenomics* **1**, 101–108.
7. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
8. Brenner, S., Johnson, M., Bridgham, J., et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*. **18**, 630–634.
9. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

10. Lashkari, D. A., DeRisi, J. L., McCusker, J. H., et al. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* **94**, 13,057–13,062.
11. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet*. **21**, 20–24.
12. Lockhart, D. J. and Winzeler, E. A. (2000) Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836.
13. 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.
14. Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo
15. SAGEGenie http://cgap.nci.nih.gov/SAGE
16. Heller, M. J. (2002) DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng*. **4**, 129–153.
17. Triche, T. J., Schofield, D., and Buckley, J. (2001) DNA microarrays in pediatric cancer. *Cancer J*. **7**, 2–15.
18. Lu, J., Lal, A., Merriman, B., Nelson, S., and Riggins, G. (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics* **84**, 631–636.
19. van Ruissen, F., Ruijter, J. M., Schaaf, G. J., et al. (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* **6**, 91.
20. Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr., et al. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*. **31**, 5676–5684.
21. Boon, K., Osorio, E. C., Greenhut, S. F., et al. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 11,547–11,548.
22. Park, P. J., Cao, Y. A., Lee, S. Y., et al. (2004) Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J. Biotechnol*. **112**, 225–245.
23. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412.

14

Clustering Analysis of SAGE Transcription Profiles Using a Poisson Approach

Haiyan Huang, Li Cai, and Wing H. Wong

Summary

To gain insights into the biological function and relevance of genes using serial analysis of gene expression (SAGE) transcription profiles, one essential method is to perform clustering analysis on genes. A successful clustering analysis depends on the use of effective distance or similarity measures. For this purpose, by considering the specific properties of SAGE technology, we modeled the SAGE data by Poisson statistics and developed two Poisson-based measures to assess similarity of gene expression profiles. By employing these two distances into a K-means clustering procedure, we further developed a software package to perform clustering analysis on SAGE data. The software implementing our Poisson-based algorithms can be downloaded from http://genome.dfci.harvard.edu/sager. Our algorithm is guaranteed to converge to a local maximum when Poisson likelihood-based measure is used. The results from simulation and experimental mouse retina data demonstrate that the Poisson-based distances are more appropriate and reliable for analyzing SAGE data compared to other commonly used distances or similarity measures.

Key Words: Clustering analysis; (Dis)similarity measures; Poisson statistics; K-means clustering; SAGE data.

1. Introduction

Serial analysis of gene expression (SAGE), an effective technique for comprehensive gene expression profiling, has been employed in studies of a wide range of biological systems ***(1–5)***. Previous efforts to develop SAGE analysis methods have been focused primarily on extracting SAGE tags and

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

identifying differences in mRNA levels between two libraries ***(2,3,6–11)***. To gain additional insights into the biological function and relevance of genes from expression data, an established strategy is to perform clustering analysis, which is to search for patterns and group transcripts with similar expression profiles. This strategy has led to the fundamental question of how to measure the (dis)similarity of gene expression across multiple SAGE libraries. An effective distance or similarity measure ***(12)***, which takes into account the underlying biology and the nature of data, would be the basis for a successful clustering analysis. Commonly used distances or similarity measures include the *Pearson correlation coefficient* and *Euclidean distance*. Pearson correlation is used to detect the shape coherence of expression curves; Euclidian distance can be used when the data are normally distributed and the magnitude of expression matters. Other measures of relationships include *likelihood-based* approaches for measuring the probabilities of clusters of genes in Gaussian mixture modeling ***(13–15)***, etc. These measures have been proven useful in microarray expression data analysis. However, SAGE data are governed by different statistics; they are generated by sampling, which results in "counts." In this regard, clustering analysis of SAGE data should involve appropriate statistical methods that consider the specific properties of SAGE data.

In one of our previous studies ***(16)***, we assumed that the tag counts follow a Poisson distribution. This is a natural assumption considering that SAGE data are generated through a random sampling technique. Based on this assumption, two Poisson-based measures were developed to assess the similarity of tag count profiles across multiple SAGE libraries ***(16)***. One measure was defined based on Chi-square statistic, which evaluates the deviation of observed tag counts from expected counts in each cluster. This method was called *PoissonC*. The other measure was based on the log-likelihood of observed tag counts, which determines the cluster membership of a transcript by its observed counts' joint probability under the expected Poisson model in each cluster. This method was called *PoissonL.* A packaged clustering program with a modified K-means procedure and with the two measures implemented is available at http://genome.dfci.harvard.edu/sager.

In this chapter, we will introduce this Poisson-based SAGE clustering method and evaluate its performance by applying it to a simulation dataset and an experimental mouse retinal SAGE dataset. These additional applications to those described in Cai et al. ***(16)*** further demonstrate the advantages of the Poisson-based measures over Pearson correlation and Euclidean distance in terms of producing clusters of more biological relevance. We also verify that the Poisson

likelihood-based clustering algorithm *PoissonL* is guaranteed to converge to a local maximum of the Poisson likelihood function for observed data.

2. Materials

1. *Software:* online web application website as well as a Linux and Microsoft Windows software are available at http://genome.dfci.harvard.edu/sager.
2. *License agreement:* the program is copyrighted by Li Cai, Haiyan Huang, and other contributors, and is free for nonprofit academic use. It can be redistributed under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or any later version. This program is distributed in the hope that it will be useful for research purpose, but **without any warranty**.
3. *Data:* the file format is a classical expression matrix, with each row representing the counts for a single tag over multiple SAGE libraries and with each column representing the counts for all tags in a single library. The packaged program prefers tab-delimited format. The specific extensions supported by the packaged program are txt, xls, wk1, wk3, wk4, mdb, fp5, 123, and dat.
4. *Minimum computer hardware requirements:* the computer used to run this program should meet at least the following requirements: (1) at least 256 MB of RAM; (2) an at least 1-GHz CPU; (3) a hard drive with at least 500 MB of free disk space; (4) Microsoft Windows 9x/NT/ME/2000/XP or any Linux operating system.

3. Methods

In the following sections, we rationalize the Poisson assumption on SAGE data and provide a detailed description on the Poisson probability model, by which two Poisson-based similarity measures were defined (*see* **Note 1**). We also verify that the introduced clustering algorithm with the likelihood-based similarity measure is guaranteed to converge to a local maximum of the Poisson likelihood function. Finally, we present the application of the Poisson-based method to a simulation dataset and a real dataset.

3.1. Poisson Assumption

In an SAGE experiment, the tag extraction is performed on a set of transcripts that are sampled from a cell or tissue. As discussed in Man et al. ***(10)***, this sampling process is approximately equivalent to randomly taking a bag of colored balls from a big box. This randomness leads to an approximate multinomial distribution for the number of transcripts of different types for tag extraction ***(17)***. Moreover, as a result of the vast amount and numerous varied types of transcripts in a cell or tissue, the selection probability of a particular

type of transcript at each draw should be very small, which suggests that the tag counts of sampled transcripts of each type can be approximately Poisson distributed.

3.2. Probability Model

The above arguments suggest a Poisson-based probability model, which can be specified by the following two assumptions.

3.2.1. Assumption 1

$Y_i(t)$, the count of tag i in library t, are independent Poisson variables with parameters $\lambda_i(t)\theta_i$, where θ_i is the expected sum of counts of tag i over all libraries (unrelated to t), $\lambda_i(t)$ is the contribution of tag i in library t to the sum (θ_i) expressed in percentage, and the sum of $\lambda_i(t)$ over all libraries equals to 1.

Assumption 1 forms the basis of the probability model. By definition, θ_i reflects the gene general expression level, $\lambda_i(t)$ describes the expression changes across libraries, and $\lambda_i(t)\theta_i$ re-distributes the tag counts according to the expression profile $[\lambda_i(t)]$ with the sum of counts across libraries kept constant. The tags with similar $\lambda_i(t)$ over t (libraries) will be grouped together, because an established strategy for finding functionally related genes is to group genes with similar expression patterns ***(18)***. This motivates Assumption 2.

3.2.2. Assumption 2

The tags in the same cluster share a common profile of $\lambda_i(t)$ over t. The common profile is denoted by $\lambda = [\lambda(1), \lambda(2), \ldots, \lambda(T)]$, where T is the total number of libraries considered. λ then represents the cluster profile.

Now, let $\boldsymbol{Y}_i = [Y_i(1), \ldots, Y_i(T)]$ denote the vector of counts of tag i across T libraries. Then, under the above two assumptions, for a cluster consisting of tags 1, 2, …, m, the joint likelihood function for $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_m$ is

$$L(\lambda, \theta \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_m} \mid \lambda, \theta_1, \ldots, \theta_m) = \prod_{i=1}^{m}\prod_{t=1}^{T} \frac{\exp(-\lambda(t)\theta_i)(\lambda(t)\theta_i)^{Y_i(t)}}{Y_i(t)!}. \quad (1)$$

The maximum likelihood estimates (MLEs) of λ and $\theta_1, \ldots, \theta_m$ are

$$\hat{\theta}_i = \sum_t Y_i(t), \text{ and } \hat{\lambda}(t) = \sum_{i=1}^{m} Y_i(t) \Big/ \sum_{i=1}^{m} \hat{\theta}_1 = \sum_{i=1}^{m} Y_i(t) \Big/ \sum_{i=1}^{m}\sum_t Y_i(t). \quad (2)$$

3.3. Two Poisson-based Similarity Measures for Clustering Tags With Similar Expression Profiles

Given a cluster consisting of tags 1, …, m, the MLEs of parameters λ and θ_i in **eq. 2** provide the expected Poisson distributions for the tag counts in each cluster. This forms the basis of the definitions of the following two measures, which evaluate how well a particular tag (gene) fits in each of the clusters.

3.3.1. Likelihood-Based Measure

It is natural to use the log-likelihood function $\log f(\boldsymbol{Y}_i|\lambda, \theta_i)$ to evaluate how well the observed counts ($\boldsymbol{Y}_i$) fit the expected Poisson distributions. The larger the log-likelihood is, the more likely the observed counts are to be generated from the expected model. For a cluster consisting of tags 1, 2, …, m, the dispersion is defined as

$$L = -\log f(Y_1, \ldots, Y_m \mid \hat{\lambda}, \hat{\theta}) = \sum_{i=1}^{m} \sum_{t=1}^{T} (\hat{\lambda}(t)\hat{\theta}_i - Y_i(t)\log(\hat{\lambda}(t)\hat{\theta}_i) + \log(Y_i(t)!)). \tag{3}$$

The optimal partition of the genes into k distinct clusters can be obtained by minimizing the cluster dispersion $L_1 + L_2 + \ldots + L_k$.

3.3.2. Chi-Square Statistic-Based Measure

The Chi-square statistic can evaluate the deviation of observed counts from expected counts in each cluster. For a cluster consisting of tags 1, 2, …, m, the dispersion can be defined as

$$D = \sum_{i=1}^{m} \sum_{t=1}^{T} (Y_i(t) - \hat{\lambda}(t)\hat{\theta}_i)^2 \Big/ (\hat{\lambda}(t)\hat{\theta}_i). \tag{4}$$

The smaller D is, the tighter the cluster is. The optimal partition of the genes into k distinct clusters can be obtained by minimizing the cluster dispersion $D_1 + D_2 + \ldots + D_k$. Using the Chi-square statistic as a similarity measure, the penalty for deviation from large expected count is smaller than that for small expected count. This is consistent with the above likelihood-based measure because the variance of a Poisson variable equals its mean.

3.4. Clustering Procedure

Using the above two measures, Cai et al. ***(16)*** modified the K-means clustering algorithm to group tags with similar count profiles. The K-means clustering procedure ***(19)*** generates clusters by specifying a desired number of clusters, say, K, and then assigns each object to one of K clusters so as to

minimize a measure of dispersion within the clusters (*see* **Note 2**). We outline the algorithm from Cai et al. ***(16)*** as follows:

1. All SAGE tags are assigned at random to K sets. Estimate initial parameters $\theta_i^{(0)}$ and $\lambda_k^{(0)} = (\lambda_k^{(0)}(1), \ldots, \lambda_k^0(T))$ for each tag and each cluster by **eq. 2**.
2. In the $(b+1)$th iteration, assign each tag i to the cluster with minimum deviation from the expected model. The deviation is measured by either $L_{i,k}^{(b)} = -\log f(Y_i \mid \lambda_k^{(b)}, \theta_i^{(b)})$ or $D_{i,k}^{(b)} = \sum_t \left(Y_i(t) - \lambda_k^{(b)}(t)\theta_i^{(b)}\right)^2 \Big/ (\lambda_k^{(b)}(t)\theta_i^{(b)})$.
3. Set new cluster centers $\lambda_k^{(b+1)}$ by **eq. 2**.
4. Repeat **step 2** until convergence.

Let $c(i)$ denote the index of the cluster that tag i is assigned to. The above algorithm aims to minimizes the within-cluster dispersion $\sum_i L_{i,c(i)}$ or $\sum_i D_{i,c(i)}$. The algorithm using the likelihood-based measure L was called *PoissonL*, and the algorithm using the Chi-square based measure D was called *PoissonC*. We want to point out that *PoissonL* is guaranteed to converge to a local maximum of the joint likelihood function for the observed data under the assumed probability model. We present the proof below.

3.4.1. Lemma 3.4.1.

Each iteration in the *PoissonL* algorithm is guaranteed to increase the likelihood for the observed data under the assumed probability model, and thus the algorithm is guaranteed to converge to a local maximum of the likelihood function.

3.4.2. Proof of Lemma 3.4.1.

Under the Poisson model described under **Subheading 3.2.**, the tag count profiles $Y_1, Y_2, \ldots, Y_N$ are assumed to be independently generated from K different joint Poisson distributions, whereas the information on which and what model generates each tag count profile is unknown. Let y_i be the cluster label for tag i, and $\Theta = (\lambda_1, \ldots, \lambda_K, \theta_1, \ldots, \theta_N)$ be the model parameters with λ_K and θ_i defined as under **Subheading 3.2.** Then, the objective is to find the Θ and y_i that maximize

$$\mathbf{L}(\Theta \mid \mathbf{Y}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_i|\Theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} f(\boldsymbol{Y}_i|\boldsymbol{\lambda}_k, \theta_i)^{\mathrm{I}(y_i=k)}, \quad (5)$$

where $\mathrm{I}(y_i = k)$ equals 1 when $y_i = k$ and 0 otherwise.

In the $(b+1)$th iteration of *PoissonL*, for $i = 1, .., N$ and $k = 1, .., K$, we estimate

$$y_i^{(b+1)} = \arg\min_k L_{i,k} = \arg\max_k f(\boldsymbol{Y}_i|\boldsymbol{\lambda}_k^{(b)}, \boldsymbol{\theta}_i^{(b)}) \quad \text{(by step 2 of the algorithm), and} \quad (6)$$

$$\boldsymbol{\Theta}^{(b+1)} = \arg\max_{\Theta} \prod_{i=1}^{N} \prod_{k=1}^{K} f(\boldsymbol{Y}_i|\boldsymbol{\lambda}_k, \theta_i)^{\mathrm{I}(y_i^{(b+1)}=k)} \quad \text{(by step 3 of the algorithm).} \quad (7)$$

$$\begin{aligned} \text{Then,} \quad & \mathbf{L}(\boldsymbol{\Theta}^{(b+1)} \mid \mathbf{Y}) = \prod_{i=1}^{N} \prod_{k=1}^{K} f(\boldsymbol{Y}_i \mid \boldsymbol{\lambda}_k^{(b+1)}, \theta_i^{(b+1)})^{\mathrm{I}(y_i^{(b+1)}=k)} \\ & \text{(by (7))} \qquad \geq \prod_{i=1}^{N} \prod_{k=1}^{K} f(\boldsymbol{Y}_i \mid \boldsymbol{\lambda}_k^{(b)}, \theta_i^{(b)})^{\mathrm{I}(y_i^{(b+1)}=k)} \\ & \text{(by (6))} \qquad \geq \prod_{i=1}^{N} \prod_{k=1}^{K} f(\boldsymbol{Y}_i \mid \boldsymbol{\lambda}_k^{(b)}, \theta_i^{(b)})^{\mathrm{I}(y_i^{(b)}=k)} = \mathbf{L}(\boldsymbol{\Theta}^{(b)} \mid \mathbf{Y}), \end{aligned} \quad (8)$$

which means that each iteration in *PoissonL* is guaranteed to increase the likelihood for the observed data, and thus the algorithm is guaranteed to converge to a local maximum.

PoissonC and *PoissonL* differs at the step of updating $y_i^{(b+1)}$. In *PoissonC*, $y_i^{(b+1)} = \arg\min_k D_{i,k}^{(b)}$, under which $f(Y_i \mid \lambda_k^{(b)}, \theta_i^{(b)})^{1(y_i^{(b+1)}=k)} \geq f(\boldsymbol{Y}_i \mid \lambda_k^{(b)}, \theta_i^{(b)})^{1(y_i^{(b)}=k)}$ and therefore $\mathrm{L}(\boldsymbol{\Theta}^{(b+1)} \mid \mathrm{Y}) \geq \mathrm{L}(\boldsymbol{\Theta}^{(b)} \mid \mathrm{Y})$ may not hold because $D_{i,k}$ is not always monotone relative to the likelihood function. The nonmonotone domain is, however, vastly small. In practice, the nonmonotone domain is often sufficiently small and negligible for the considered dataset such that *PoissonC* agrees with *PoissonL* and converges to a local maximum. One big advantage of *PoissonC* compared to *PoissonL* is that it runs much faster based on the current version of program (*see* **Note 3**).

PoissonL is actually a specific version of Classification EM algorithm (CEM) ***(20)***. The objective likelihood function of CEM under the mixture Poisson model is

$$\mathbf{L}_{\mathrm{CEM}}(\boldsymbol{\Theta} \mid \mathbf{Y}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left(f(\mathbf{Y}_i \mid \Theta)^{\mathrm{I}(y_i=k)} f(\mathbf{y}_i = k \mid \boldsymbol{\Theta}) \right), \quad (9)$$

which is equivalent to **eq. 5** when the prior conditional probability of y_i given Θ is uniform (*see* **Note 4**).

3.5. Implementation

PoissonL and *PoissonC* are implemented in both C++ and Java. The implementation in C++ is based on the open source code of the C clustering Library provided by de Hoon et al. ***(21)*** (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) (*see* **Note 5**). We developed a web-based application as well as Microsoft Windows and Linux versions of software to perform the clustering analysis. The software is available at http://genome.dfci.harvard.edu/sager.

3.6. Examples

Two examples are presented here to demonstrate the advantages of Poisson-based measures over other commonly used distance or similarity measures in analyzing SAGE or Poisson-like data. Because these examples are independent of the ones shown in Cai et al. ***(16)***, they can serve as an additional validation of the Poisson-based measures.

3.6.1. Example I: Clustering Results of Simulation Data

Data. The distributions used to generate the simulation dataset are described in **Table 1**. The simulation dataset consists of 46 vectors of dimension 5 with components independently generated from different Normal distributions. The mean (μ) and variance (σ^2) parameters of the normal distributions are constrained by $\sigma^2 = 3\,\mu$. This application evaluates the performance of our method on data with Poisson-like properties: variance increases with mean. Success in this dataset would shed light on more broad applications of our method.

In our simulation dataset, the 46 vectors belong to six groups (named A, B, C, D, E, and F) according to the Normal distributions from which they are generated. The six groups are of size 3, 6, 6, 9, 7, and 15, respectively. For comparison, we applied *PoissonC* together with *Eucli* (classical K-means clustering algorithm

Table 1
5-Dim Simulation Dataset With Normal Distributions $\sigma^2 = 3\mu$

Group ID		Mean parameters of the normal distributions (μ)				
Group A	a1 ~ a3	1	1	1	15	150
Group B	b1 ~ b6	15	1	1	1	150
Group C	c1 ~ c4	10	30	30	60	10
	c5 ~ c6	100	300	300	600	100
Group D	d1 ~ d7	200	70	70	10	10
	d8 ~ d9	2000	700	700	100	100
Group E	e1 ~ e5	210	120	10	10	10
	e6 ~ e7	2100	1200	100	100	100
Group F	f1 ~ f3	5	50	5	5	5
	f4 ~ f6	5	75	5	5	5
	f7 ~ f9	5	100	5	5	5
	f10 ~ f11	50	500	50	50	50
	f12 ~ f13	50	750	50	50	50
	f14 ~ f15	50	1000	50	50	50

using Euclidian distance) and *PearsonC* (K-means clustering procedure using Pearson correlation as similarity measure) to the simulated data. The clustering results from different methods are shown in **Fig. 1**. The simulation data is available at http://www.stat.berkeley.edu/users/hhuang/SAGE.html.

Results. In **Fig. 1**, only *PoissonC* has clustered the vectors perfectly into six groups. All of the other methods fail to correctly separate the vectors from Group A and Group B. *Eucli* works the worst when it is applied to unnormalized data. It fails to identify any of the six clusters. This is because Euclidian distance can be overly sensitive to the magnitude of changes. To reduce the magnitude effects, we further apply *Eucli* to the rescaled data. The rescaling is performed so that the sum of the components within each vector is set the same. The clustering result of *Eucli* on rescaled data is clearly better than the

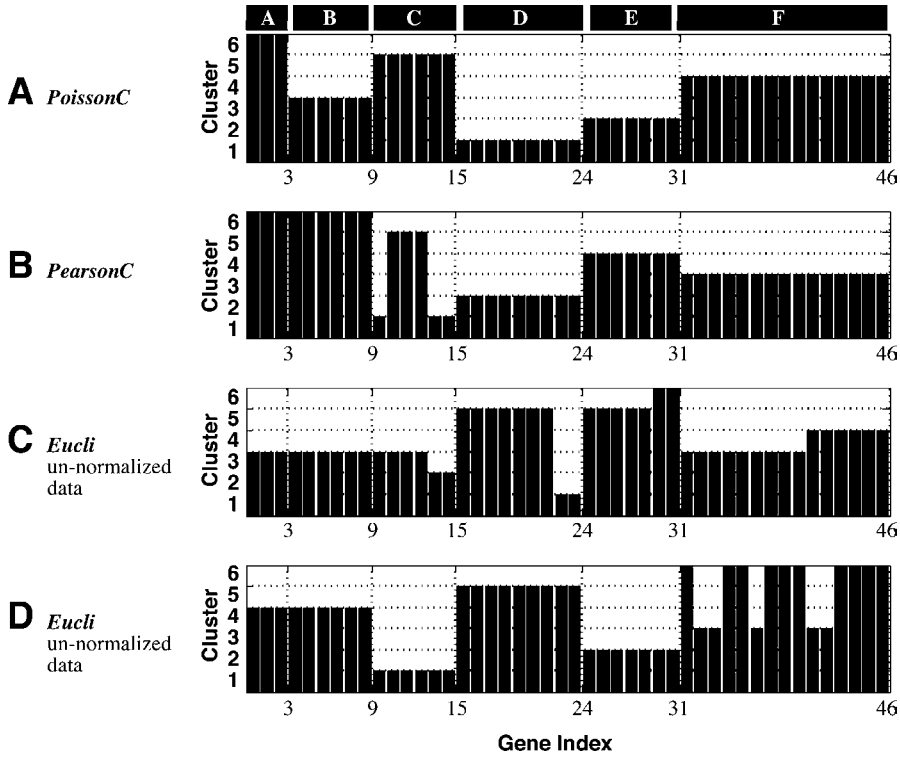


Fig. 1. Graphs of clustering results for simulation data. Horizontal axis represents the index of the 46 vectors, which belong to six groups (named A, B, C, D, E, and F) that are marked at the top of the figure. Vertical axis represents the index of the cluster that each vector has been assigned to by each algorithm.

result on unnormalized data. Groups C, D, and E have been correctly identified (*see* **Note 6**).

We perform an additional 100 replications of the above simulation. *PoissonC* correctly clusters 34 of the 100 replicate datasets. *Eucli*, on rescaled data, correctly clusters 2 of the 100 datasets whereas *PearsonC* or *Eucli*, on unnormalized data, never generates correct clusters.

We also want to point out that there is a small error in the simulation results presented in Table 1 and Fig. 1 of Cai et al. ***(16)***. For the data in Table 1, Fig. 1 reported a perfect clustering result by *PoissonC*, which is not correct. But the conclusion made from that example that *PoissonC* is superior to other methods is still valid because *PoissonC* has only wrongly clustered one tag.

3.6.2. Example II: Clustering Results of Experimental SAGE Data

For further validation, we apply *PoissonC, PearsonC*, and *Eucli* to a set of mouse retinal SAGE libraries.

Data. The raw mouse retinal data consists of 10 SAGE libraries (38,818 unique tags with tag counts ≥ 2) from developing retina taken at 2-d intervals, ranging from embryonic to postnatal and adult ***(16,22)***. One thousand four hundred sixty-seven of the 38,818 tags with counts ≥ 20 in at least one of the 10 libraries are selected (*see* **Note 7**). To effectively compare the clustering algorithms, a subset of 153 SAGE tags with known biological functions are further selected (*see* **Note 8**). These 153 tags fall into five clusters based on their biological function(s) (*see* **Table 2a**). One hundred twenty-five of these genes are developmental genes, which can be further grouped into four clusters by their expressions at different developmental stages. The other 28 genes are unrelated to the mouse retina development. This dataset is available at http://www.stat.berkeley.edu/users/hhuang/SAGE.html.

Results. PoissonC, PearsonC, and *Eucli* are applied to group these 153 tags into five clusters. Results show that the performance of *PoissonC* is superior to other methods (*see* **Table 2b**). We should also note that *PoissonC* is only

Table 2a
Functional Categorization of the 153 Mouse Retinal Tags (125 Developmental Genes; 28 Nondevelopmental Genes)

	Function Groups					
	Early I	Early II	Late I	Late II	Non-dev.	Total
Number of tags	32	34	32	27	28	153

Table 2b
Comparison of Algorithms on 153 Tags

Algorithm	# of tags in incorrect clusters	% of tags in incorrect clusters
PoissonC	22	14.4
Eucli on normalized data	36	23.5
PearsonC	26	17.0
Eucli	NA	NA

Clusters generated by *Eucli* were too messy.

slightly better than *PearsonC* in this application because the shapes of the gene expression curves are quite different from each other among these five clusters and the Pearson correlation can powerfully detect the shape coherence of curves.

Acknowledgments

The method described in this chapter is based on the original research paper published in Genome Biology ***(16)***. We thank Kyungpil Kim for help in generating the figure and tables.

4. Notes

1. The main advantage of the described method is that the newly designed measures consider both the magnitude and shape when comparing the expression patterns (λ represents the shape and θ represents the magnitude in our model), whereas Euclidian distance is focused only on the magnitude of changes and Pearson correlation is overly sensitive to the shape of the curve.
2. An unsolved issue in K-means clustering analysis is the estimation of K, the number of clusters. If K is unknown, starting with arbitrary, random K is a relatively poor method. Hartigan proposed a stage-wise method to determine the K value ***(19)***. However, when sporadic points are present in the dataset, Hartigan's method may fail. A recently introduced method, TightCluster ***(23)***, partially solves this problem by using a resampling scheme to sequentially attain tight and stable clusters in the order of decreasing stability. The Poisson based measures can be implemented in the TightCluster program to apply the TightCluster method to SAGE data.
3. *PoissonL* and *PoissonC* performed similarly when they were applied to many small simulation and experimental data sets. For large datasets, *PoissonC* should be more practical, as the current version of *PoissonL* (installed in the software package) is too slow. There is still much room for improving the *PoissonL* algorithm.

4. We can also derive an EM algorithm for fitting the mixture Poisson model. The associated objective likelihood is

$$\mathbf{L}_{\mathrm{EM}}(\Theta|\mathbf{Y}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_i|\boldsymbol{\Theta}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} f(\boldsymbol{Y}_i|\boldsymbol{\Theta}) f(y_i = k|\boldsymbol{\Theta}) \right). \tag{10}$$

The E-step and M-step of the algorithm can be described as follows:
E-step: with the estimated $\boldsymbol{\Theta}^{(b)} = (\boldsymbol{\lambda}_1^{(b)}, \ldots, \boldsymbol{\lambda}_K^{(b)}, \theta_1^{(b)}, \ldots, \theta_N^{(b)})$, compute

$$\begin{aligned} & Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(b)}) \\ & = E_{y_1,\ldots,y_N|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_N,\boldsymbol{\Theta}^{(b)}} [\log f(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N, y_1, \ldots, y_N|\boldsymbol{\Theta})] \\ & = \sum_{y_1,\ldots,y_N} \left(\log \prod_{i=1}^{N} f(\boldsymbol{Y}_i|\boldsymbol{\Theta}) \right) f(y_1, \ldots, y_N|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N, \boldsymbol{\Theta}^{(b)}) \\ & = \sum_{y_1,\ldots,y_N} \left(\log \prod_{i=1}^{N} \left(\sum_{k=1}^{K} f(Y_i|\boldsymbol{\Theta}) f(\boldsymbol{y}_i = k|\boldsymbol{\Theta}) \right) \right) \left(\prod_{i=1}^{N} f(y_i|\boldsymbol{Y}_i, \boldsymbol{\Theta}^{(b)}) \right) \end{aligned} \tag{11}$$

M-step: find $\Theta^{(b+1)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(b)})$.
Clearly, in the above EM algorithm, the objective likelihood function and therefore the optimal clustering results depend on the prior conditional probability of y_i given Θ. Preliminary simulation comparisons among *PoissonL, PoissonC*, and the EM algorithm show that they perform similarly. Further comparisons of these algorithms are ongoing.
5. The new measures were employed into a K-means clustering procedure to perform the analysis. The algorithm used for iteratively updating cluster assignments is an algorithm implemented in the C clustering library, which is publicly available ***(21)***. The algorithm terminates when no further reassignments take place. Because the convergent results of this algorithm are quite sensitive to the initial cluster assignments, usually, the algorithm should be run on many different initials to obtain an optimal result. The within-cluster dispersion should better be recorded to compare the results.
6. When the users are not confident about whether the data are Poisson-like or not, a good choice could be *Eucli* (K-means algorithm using Euclidian distance). Our experience tells that *Eucli* is quite stable and reliable when it is applied to data that are appropriately postnormalized according to the clustering purpose, i.e., the data can be rescaled to reduce the effects of magnitude if only the shape of expression pattern determines the clustering. Good measurement methods should consider both magnitude and shape of the expression patterns.
7. For clustering analysis, tags with only one count are usually excluded from analysis due to sequencing error problem. To select the potential most biologically relevant genes, tags with less than 2–10 counts can be excluded depending on how large the SAGE libraries are and how many total number of tags is intended to analyze.
8. Annotation of SAGE tags is through SAGEtag to UniGene mapping ***(24)***. The mapping is based on "SAGEmap_tag_ug-rel.Z" provided by the National center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/), which contains all annotated SAGE tags mapping to UniGene clusters. However, there

are many ambiguities on the SAGE tag annotation. There are tag sequencing errors ***(25)***, and also the mapping between tags and genes can be nonunique. In one planned project, we propose to reduce this error by inferring the real expression level of genes from "weighted" counts of all mapped tags, where the weights can be determined by the available mapping quality information. An EM algorithm is feasible for this task.

References

1. Blackshaw, S., Fraioli, R. E., Furukawa. T., and Cepko, C. L. (2001) Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell* **107**, 579–589.
2. Zhang, L., Zhou, W., Velculescu, V. E., et al. (1997) Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272.
3. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.
4. Buckhaults, P., Zhang, Z., Chen, Y. C., et al. (2003) Identifying tumor origin using a gene expression-based classification map. *Cancer Res.* **63**, 4144–4149.
5. Porter, D., Weremowicz, S., Chin, K., et al. (2003) A neural survival factor is a candidate oncogene in breast cancer. *Proc Natl Acad Sci USA*. **100**, 10,931–10,936.
6. Margulies, E. H. and Innis, J. W. (2000) eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics* **16**, 650–651.
7. van Ruissen, F., Jansen, B. J., de Jongh, G. J., van Vlijmen-Willems, I. M., and Schalkwijk, J. (2002) Differential gene expression in premalignant human epidermis revealed by cluster analysis of serial analysis of gene expression (SAGE) libraries. *FASEB J*. **16**, 246–248.
8. Audic, S. and Claverie, J. M. (1997) The significance of digital gene expression profiles. *Genome Res*. **7**, 986–995.
9. Madden, S. L., Galella, E. A., Zhu, J., Bertelsen, A. H., and Beaudry, G. A. (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
10. Man, M. Z., Wang, X., and Wang, Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*. **16**, 953–959.
11. Blackshaw, S., Kuo, W. P., Park, P. J., et al. (2003) MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues. *Genome Biol.* **4**, R17.
12. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet*. **2**, 418–427.
13. Fraley, C. (1998) Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing* **20**, 270–281.

14. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001) Model-based clustering and data transformation for gene expression data. *Bioinformatics* **17**, 977–987.
15. Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**, 611–631.
16. Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C. L., and Wong, W. H. (2004) Clustering analysis of SAGE data using a Poisson approach. *Genome Biol.* **5**, R51.
17. Ewens, W. J. and Grant, G. R. (2001) *Statistical Methods in Bioinformatics*. Springer Verlag, Germany.
18. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14,863–14,868.
19. Hartigan, J. (1975) *Clustering Algorithms*. Wiley, New York.
20. Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* **14**, 315–332.
21. de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004) Open source clustering software. *Bioinformatics* **20**, 1453–1454.
22. Blackshaw, S., Harpavat, S., Trimarchi, J., et al. (2004) Genomic analysis of mouse retinal development. *PLoS Biology* **2**, E247.
23. Tseng, G. C. and Wong, W. H. (2004) A resampling method for tight clustering: with an application to microarray analysis. *Biometrics* **61**, 10–16.
24. Lash, A. E., Tolstoshev, C. M., Wagner, L., et al. (2000) SAGEmap: a public gene expression resource. *Genome Res*. **10**, 1051–1060.
25. Beissbarth, T., Hyde, L., Smyth, G. K., et al. (2004) Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics*. **4(Suppl 20)** 1:I31–I39.

15

Identifying Nonspecific SAGE Tags by Context of Gene Expression

Xijin Ge and San Ming Wang

Summary

Many serial analysis of gene expression (SAGE) tags can be matched to multiple genes, leading to difficulty in SAGE data interpretation and analysis. As only a subset of genes in the human genome are transcribed in a certain type of tissue/cell, we used microarray expression data from different tissue types to define contexts of gene expression and to annotate SAGE tags collected from the same or similar tissue sources. To predict the original transcript contributing a nonspecific SAGE tag collected from a particular tissue, we ranked the corresponding genes by their expression levels determined by microarray. We developed a tissue-specific SAGE tag annotation database based on microarray data collected from 73 normal human tissues and 18 cancer tissues and cell lines. The database can be queried online at: http://www.basic.northwestern.edu/SAGE/. The accuracy of this database was confirmed by experimental data.

Key Words: SAGE; microarray; tag annotation; non-specific SAGE tag; tissue-specificity.

1. Introduction

Reference database plays a key role in mapping serial analysis of gene expression (SAGE) tags to genes. The widely used SAGEmap ***(1)*** and SAGE Genie ***(2)*** are constructed by comparing the tag sequences to expressed sequences deposited in public expression databases. However, a significant portion of conventional 14-bp-long tags is nonspecific ***(3)***, and could be originated from multiple genes. This leads to difficulties in downstream functional analysis, and limits the usefulness of SAGE data in transcriptome study.

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

Public mRNA and expressed sequence tag (EST) sequences used to construct tag-to-gene mappings are derived from diverse sources, including those from various tissues at various developmental stages, or under different pathological conditions. Considering that many genes are expressed only or mainly in a specific tissue type but not or rarely in the others, it is possible to improve the tag-to-gene mapping by putting short SAGE tags into a context of tissue origin. The specificity of SAGE tags could be increased if we only compare tag sequence with transcripts expressed in the corresponding tissue type. However, it is difficult to directly parse dbEST to classify these sequences as a result of the diversity of tissue sources and the lack of well-defined format and vocabulary.

DNA microarray has been used for systematic studies on tissue-specific gene expression ***(4–6)***, based on longer probes (>24 bp) specifically designed to detect known transcripts. In a tissue-specific manner, such microarray data could be employed to predict the most likely contributing gene for nonspecific SAGE tags. We developed a microarray data-based, tissue-specific SAGE tag-to-gene mapping methodology that will enable the unique mapping of thousands of ubiquitous tags ***(7)***.

2. Construction of the Database

2.1. Tag-to-Gene Mapping

SAGEmap "full" and "reliable" mappings are downloaded from the National Center for Biotechnology Information (NCBI) (human, *Nla*III, build 182, http://www.ncbi.nlm.nih.gov/SAGE/index.cgi?cmd=mappings).

2.2. DNA Microarray Data for Normal Tissues

Su et al. conducted a systematic microarray study of gene expression in human tissues (***6***; http://symatlas.gnf.org). Their dataset covered 73 normal tissue types and 6 cancerous tissue types or cell lines. Each tissue was represented by two replicates of pooled RNA samples. The expression information was collected by using the combination of HG-U133A array (Affymetrix, Santa Clara, CA) and a custom array designed to detect additional transcripts mainly based on *in silico* predictions. Raw microarray (Affymetrix. cel) files are provided by the authors. In addition, microarray data of 12 common cancers reported in **ref. *8*** was downloaded from http://carrier.gnf.org/welsh/epican/index.htm.

2.3. DNA Microarray Data Processing

1. Raw microarray image (.CEL) data were processed using Microarray Suite (MAS) 5.0 software to produce expression score ("signal") and a detection *P*-value.

2. Starting from probe annotation files downloaded from the Affymetrix web site (http://www.affymetrix.com), we extracted the target sequence IDs used for probe design.
3. Annotation is performed by searching for this GenBank accession number in the latest version of UniGene. We linked 22,283 probe sets to 13,263 UniGene clusters (Build #182).
4. For the custom array, a BLAST search was performed against representative sequences in UniGene database. We mapped 3606 target sequences to 2782 UniGene clusters with a cutoff E-value of 1×10^{-50}. Together, 15,045 UniGene clusters were identified from the microarray data covering 79 human tissue and cell types.
5. Microarray data from 12 common cancer tissues collected by Affymetrix HG-U95 array *(8)* was also incorporated. A total of 12,532 probe sets was matched to 7438 UniGene clusters.

2.4. Construction of Tissue-Specific Annotation Database

The process is illustrated in **Fig. 1**.

1. Calibrate the DNA microarray data. The averaged expression score from two replicates was applied to determine the genes expressed in each tissue type. If the confidence in expression levels of both replicates was low as indicated by the detection *P*-value (>0.06), the gene was not considered expressed in this tissue. For genes represented by more than one probes, all probe information was included to enhance specificity.

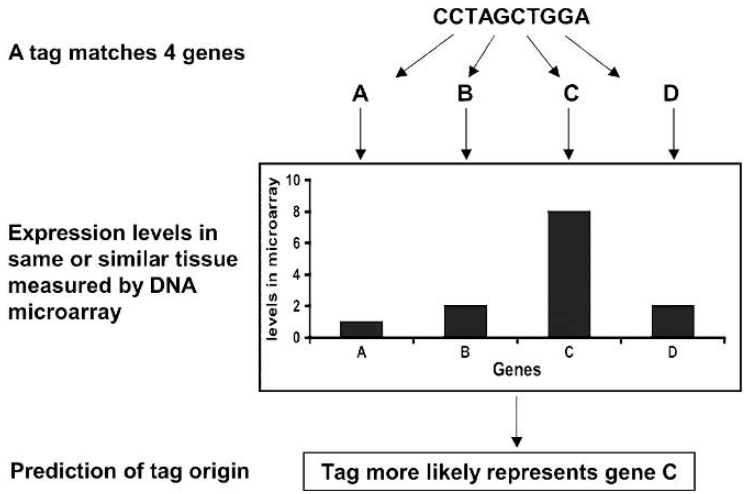


Fig. 1. The process for annotating nonspecific serial analysis of gene expression tags based on DNA microarray data.

2. Identify the SAGE tags matched to multiple UniGene clusters in the SAGEmap "full" and "reliable" database.
3. The UniGene clusters matched by the same SAGE tag are ranked according to their expression levels in microarray data. As highly expressed genes are more probable contributors of the tag, a probability score was calculated for each UniGene cluster by dividing its expression level by the sum of expression levels of all clusters matching the same tag. Suppose there are N clusters matching to the same tag and g_i $(i = 1 \ldots N)$ is the expression score of the i-th cluster; the probability score p_i for the i-th cluster is calculated by $p_i = g_i \Big/ \sum_1^N g_j$.
4. Such predictions for 91 types of tissues/cell lines are incorporated into a MySQL database and a Perl script is constructed to enable online query of this database (http://www.basic.northwestern.edu/SAGE/). For a list of SAGE tags, this script first searches against a list of uniquely mapped tags that are, according to SAGEmap, "full" or "reliable," and the remaining tags are then queried from a table containing predicted tag-to-gene mapping in the selected tissue. The database is updated periodically.

2.5. Experimental Confirmation of the Predicted Genes

The human CD34+ hematopoietic cell SAGE library ***(9)*** was downloaded from the gene expression omnibus (GEO) database. A set of 54 nonspecific tags is chosen for experimental confirmation using the generation of longer complementary DNA (cDNA) for the gene identification method ***(10)***. In the GLGI reaction, a SAGE tag was used as the sense primer, the tail sequence located at the 3′ end of cDNA incorporated during cDNA synthesis was used as the antisense primer, and the CD34+ cDNA was used as the template for PCR amplification (*see* **ref. *10*** for more details). The amplified 3′ cDNA was cloned and sequenced, and their corresponding UniGene clusters were determined by BLAST against the UniGene database. When comparing predicted identifications of a group of 54 multiple-matching tags with those that obtained through the GLGI method, we observed agreement in 49 (or 90 %) of the tags.

3. Database Query

The tag-to-gene mapping database can be accessed through an online query interface at http://www.basic.northwestern.edu/SAGE/. Users first select a tissue type that resembles that of user's SAGE library. Multiple 10-bp tags (without CATG sites) can be posted in one query. The output of annotation reports the query SAGE tag, Affymetrix microarray probe ID, tissue type, ranking and score of the annotation, and the UniGene cluster ID, GenBank accession number, full name, symbol, and Locus Link of the annotated gene.

When the exact tissue type used for SAGE library construction is not included in the database, a closely related tissue type could be used for the annotation as a common set of genes might be expressed between these tissue types. Tissue selections can be made based on common developmental origins of endoderm, mesoderm, or ectoderm.

The current database of tag annotation is also helpful in the data mining of existing SAGE libraries stored in public gene expression databases. It is now possible to predict the identity of the nonspecific tags that are differentially expressed among libraries and perform functional analysis such as searching for functionally overrepresented gene ontology (GO) terms. Although experimental confirmation using methods like GLGI will be needed, our database makes it possible to analyze hundreds or thousands of nonspecific SAGE tags that are related to certain diseases or biological activity.

Acknowledgments

The authors are indebted to Qingfa Wu and Yong-chul Jung for their help with experimental confirmation using the GLGI method, and Warren Kibbe for help with construction of web interface. We thank John B. Hogenesch for sharing DNA microarray data, Hui Dong, Wendy S. Rubinstein, and Yeong C. Kim for stimulating discussions, and Jiang Fu for comments on the presentation of the manuscript. This work was supported by the National Institutes of Health and the Daniel F. and Ada L. Rice Foundation.

References

1. Lash, A. E., Tolstoshev, C. M., Wagner, L., et al. (2000) SAGEmap: a public gene expression resource. *Genome Res*. **10,** 1051–1060.
2. Boon, K., Osorio, E. C., Greenhut, S. F., et al. (2002) An anatomy of normal and malignant gene expression. *PNAS* **99**, 11,287–11,292.
3. Lee, S., Clark, T., Chen, J., et al. (2002) Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics* **79**, 598–602.
4. Hsiao L. L., Dangond, F., Yoshida, T., et al. (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7,** 97–104.
5. Saito-Hisaminato, A., Katagiri, T., Kakiuchi, S., Nakamura, T., Tsunoda, T., and Nakamura, Y. (2002) Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res.* **9**, 35–45.
6. Su, A. I., Wiltshire, T., Batalov, S., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS* **101**, 6062–6067.
7. Ge, X. J., Jung, Y. C., Wu, Q. F., Kibbe, W. A., and Wang, S. M. (2006) Annotating non-specific SAGE tags with microarray data. *Genomics* **87,** 173–180.

8. Su, A. I., Welsh, J. B., Sapinoso, L. M., et al. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*. **61**, 7388–7393.
9. Zhou, G., Chen, J., Lee, S., Clark, T., Rowley, J. D., and Wang, S. M. (2001) The pattern of gene expression in human CD34(+) stem/progenitor cells. *PNAS* **98**, 13,966–13,971.
10. Chen, J. J., Rowley, J. D., and Wang, S. M. (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. *PNAS* **97**, 349–353.

Index