METHODS IN MOLECULAR BIOLOGY™ 311

Pharmacogenomics

Methods and Protocols

Edited by

Federico Innocenti, MD, PhD

HUMANA PRESS

Pharmacogenomics

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

323. Arabidopsis Protocols, *Second Edition,* edited by *Julio Salinas and Jose J. Sanchez-Serrano, 2006*

322. *Xenopus* Protocols: *Cell Biology and Signal Transduction,* edited by *X. Johné Liu, 2006*

321. Microfluidic Techniques: *Reviews and Protocols,* edited by *Shelley D. Minteer, 2006*

320. Cytochrome P450 Protocols, *Second Edition,* edited by *Ian R. Phillips and Elizabeth A. Shephard, 2006*

319. Cell Imaging Techniques, *Methods and Protocols,* edited by *Douglas J. Taatjes and Brooke T. Mossman, 2006*

318. Plant Cell Culture Protocols, *Second Edition,* edited by *Victor M. Loyola-Vargas and Felipe Vázquez-Flota, 2005*

317. Differential Display Methods and Protocols, *Second Edition,* edited by *Peng Liang, Jonathan Meade, and Arthur B. Pardee, 2005*

316. Bioinformatics and Drug Discovery, edited by *Richard S. Larson, 2005*

315. Mast Cells: *Methods and Protocols,* edited by *Guha Krishnaswamy and David S. Chi, 2005*

314. DNA Repair Protocols: *Mammalian Systems, Second Edition,* edited by *Daryl S. Henderson, 2005*

313. Yeast Protocols: *Second Edition,* edited by *Wei Xiao, 2005*

312. Calcium Signaling Protocols: *Second Edition,* edited by *David G. Lambert, 2005*

311. Pharmacogenomics: *Methods and Protocols,* edited by *Federico Innocenti, 2005*

310. Chemical Genomics: *Reviews and Protocols,* edited by *Edward D. Zanders, 2005*

309. RNA Silencing: *Methods and Protocols,* edited by *Gordon Carmichael, 2005*

308. Therapeutic Proteins: *Methods and Protocols,* edited by *C. Mark Smales and David C. James, 2005*

307. Phosphodiesterase Methods and Protocols, edited by *Claire Lugnier, 2005*

306. Receptor Binding Techniques: *Second Edition,* edited by *Anthony P. Davenport, 2005*

305. Protein–Ligand Interactions: *Methods and Protocols,* edited by *G. Ulrich Nienhaus, 2005*

304. Human Retrovirus Protocols: *Virology and Molecular Biology,* edited by *Tuofu Zhu, 2005*

303. NanoBiotechnology Protocols, edited by *Sandra J. Rosenthal and David W. Wright, 2005*

302. Handbook of ELISPOT: Methods and Protocols, edited by *Alexander E. Kalyuzhny, 2005*

301. Ubiquitin–Proteasome Protocols, edited by *Cam Patterson and Douglas M. Cyr, 2005*

300. Protein Nanotechnology: *Protocols, Instrumentation, and Applications,* edited by *Tuan Vo-Dinh, 2005*

299. Amyloid Proteins: *Methods and Protocols,* edited by *Einar M. Sigurdsson, 2005*

298. Peptide Synthesis and Application, edited by *John Howl, 2005*

297. Forensic DNA Typing Protocols, edited by *Angel Carracedo, 2005*

296. Cell Cycle Protocols, edited by *Tim Humphrey and Gavin Brooks, 2005*

295. Immunochemical Protocols, *Third Edition,* edited by *Robert Burns, 2005*

294. Cell Migration: *Developmental Methods and Protocols,* edited by *Jun-Lin Guan, 2005*

293. Laser Capture Microdissection: *Methods and Protocols,* edited by *Graeme I. Murray and Stephanie Curran, 2005*

292. DNA Viruses: *Methods and Protocols,* edited by *Paul M. Lieberman, 2005*

291. Molecular Toxicology Protocols, edited by *Phouthone Keohavong and Stephen G. Grant, 2005*

290. Basic Cell Culture, *Third Edition,* edited by *Cheryl D. Helgason and Cindy Miller, 2005*

289. Epidermal Cells, *Methods and Applications,* edited by *Kursad Turksen, 2005*

288. Oligonucleotide Synthesis, *Methods and Applications,* edited by *Piet Herdewijn, 2005*

287. Epigenetics Protocols, edited by *Trygve O. Tollefsbol, 2004*

286. Transgenic Plants: *Methods and Protocols,* edited by *Leandro Peña, 2005*

285. Cell Cycle Control and Dysregulation Protocols: *Cyclins, Cyclin-Dependent Kinases, and Other Factors,* edited by *Antonio Giordano and Gaetano Romano, 2004*

284. Signal Transduction Protocols, *Second Edition,* edited by *Robert C. Dickson and Michael D. Mendenhall, 2004*

283. Bioconjugation Protocols, edited by *Christof M. Niemeyer, 2004*

282. Apoptosis Methods and Protocols, edited by *Hugh J. M. Brady, 2004*

281. Checkpoint Controls and Cancer, Volume 2: *Activation and Regulation Protocols,* edited by *Axel H. Schönthal, 2004*

280. Checkpoint Controls and Cancer, Volume 1: *Reviews and Model Systems,* edited by *Axel H. Schönthal, 2004*

279. Nitric Oxide Protocols, *Second Edition,* edited by *Aviv Hassid, 2004*

278. Protein NMR Techniques, *Second Edition,* edited by *A. Kristina Downing, 2004*

277. Trinucleotide Repeat Protocols, edited by *Yoshinori Kohwi, 2004*

METHODS IN MOLECULAR BIOLOGY™

Pharmacogenomics

Methods and Protocols

Edited by

Federico Innocenti, MD, PhD

Committee on Clinical Pharmacology and Pharmacogenomics
Section of Hematology/Oncology, Department of Medicine
University of Chicago, Chicago, IL

HUMANA PRESS TOTOWA, NEW JERSEY

999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

This publication is printed on acid-free paper. ∞
ANSI Z39.48-1984 (American Standards Institute)

Permanence of Paper for Printed Library Materials.

Production Editor: Jennifer Hackworth

Cover design by Patricia F. Cleary

Cover illustration: Figure 2 from Chapter 4, “Genome-Wide Analysis of Allele-Specific Gene Expression Using Oligo Microarrays," by Maxwell P. Lee.

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel: 973-256-1699; Fax: 973-256-8341; E-mail: orders@humanapr.com; or visit our Website: www.humanapress.com

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1

ISSN 1064-3745

E-ISBN 1-59259-957-5

Library of Congress Cataloging-in-Publication Data

Pharmacogenomics : methods and applications / edited by Federico Innocenti.

p. ; cm. -- (Methods in molecular biology, ISSN 1064-3745 ;311)

Includes bibliographical references and index.ISBN 1-58829-440-4 (alk. paper)

1. Pharmacogenomics. I. Innocenti, Federico.II. Series: Methods in molecular biology (Clifton, N.J.); v. 311.[DNLM: 1. Pharmacogenetics--methods.
2. Genomics--methods.
3. Variation (Genetics) QV 38 P53193 2005] RM301.3.G45P427 2005 615'.7--dc22

2005010328

Preface

For the first time in the published literature, *Pharmacogenomics: Methods and Protocols* describes the newest and most commonly adopted technologies in the field of pharmacogenomics, providing guidance for investigators in the selection and experimental application of such technologies. Many of the contributors to this book are leading experts in the field. Using the extensive information provided on materials and methods, investigators will be able to easily reproduce each technique in their laboratories. Moreover, this book highlights problems that might be encountered in performing specific techniques and describes how to identify and overcome them. Pharmacologists, geneticists, molecular biologists, and physicians in academic institutions and the biotechnology and pharmaceutical industries will find *Pharmacogenomics: Methods and Protocols* an essential reference.

Pharmacogenomics exists at the intersection of pharmacology and genomics. It aims to study the genetic basis of interpatient variability in response to drug therapy. Pharmacogenomics holds the promise that drugs may eventually be tailor-made for individuals and adapted to each person's genetic makeup. Environment, diet, age, lifestyle, and disease state can all influence a patient's response to medicines, but understanding an individual's genetic makeup is thought to be the key to creating personalized drugs with greater efficacy and safety. Pharmacogenomics combines traditional pharmaceutical sciences with annotated knowledge of genes, proteins, and single nucleotide polymorphisms. Various technologies are currently available and researchers must be capable of choosing the technology suitable for their purposes.

After an introductory chapter about the history of pharmacogenomics and its current status, *Pharmacogenomics: Methods and Protocols* is divided in three parts. Part I comprises the methodologies for assessing the functional consequences of a certain polymorphism. Part II describes the variety of genotyping platforms currently available. Part III ends the book with two chapters devoted to the management of pharmacogenomic information.

A large amount of data about the pattern of human genomic variation has been provided by the Human Genome Project and is now publicly available. However, the functional consequences of SNPs and haplotypes are, for the most part, unknown, and current research efforts are oriented toward the elucidation of the genetic basis of changes in function of expression of the coded protein. Chapter 2 reports the classical method of transient expression combined with site-directed mutagenesis to study the functional effect of naturally occurring

variants in the UDP-glucuronosyltransferase 1A1 gene. Chapters 3–5 describe newer methodologies recently introduced to evaluate differences in gene expression between two genotypes/haplotypes. The allele-specific differential expression method described in Chapter 3 circumvents the analytic problems of confounding variation arising from environmental or physiologic factors during the analysis of subtle differences in expression between two different alleles. Chapter 4 provides a method for performing both genotyping and allele-specific gene expression for hundreds of genes using a chip system. Finally, it is crucial to take into account the haplotype structure of multiple variants when functional assays of single variants are performed. The HaploChIP is an in vivo cell-based assay that allows screening of haplotypes for differences in relative gene expression and is described in Chapter 5.

Part III deals with genotyping techniques. Chapters 6–13 present a wide variety of methodologies, platforms, and chemistries for genotyping. This section provides an understanding of the factors influencing the efficiency of different genotyping methods and the priorities required of different study designs. Readers will find technical information on several different types of assays, including denaturing high-performance liquid chromatography, pyrosequencing, kinetic-fluorescence detection assay, mass spectrometry, and TaqMan assay for insertion/deletions. Moreover, Part III describes the recent application in which genetic variation is surveyed in DNA pools from individuals enrolled in large studies.

The integration of genome-information management systems with patient clinical data sets is the key needed to achieve personalized medicine. Disparate data sources, including public or proprietary biology databases and laboratory and clinical information management systems, pose significant challenges in converting this information into clinically applicable knowledge. In Part IV, Chapter 14 describes PharmGKB, a registration-free interactive tool displaying genotype, molecular, and clinical primary data integrated with literature information and links to external sources. Finally, Chapter 15 gives an overview of the main technologies needed for the management of pharmacogenomic information.

I am extremely grateful to all the authors for their excellent contributions making this book a comprehensive and up-to-date resource for investigators in pharmacogenomics.

Federico Innocenti, MD, PhD

Contents

Contributors

Russ B. Altman • *Department of Genetics, Stanford University School of Medicine, Stanford, CA*

William R. Beggs • *Division of Molecular Biology, Coriell Institute for Medical Research, Camden, NJ*

Patrick K. Bender • *Division of Molecular Biology, Coriell Institute for Medical Research, Camden, NJ*

Lee Butcher • *SGDP Centre, Institute of Psychiatry, London, UK*

Pei Xian Chen • *Department of Human Genetics, The University of Chicago, Chicago, IL*

Ian Craig • *SGDP Centre, Institute of Psychiatry, London, UK*

Soma Das • *Department of Human Genetics, The University of Chicago, Chicago, IL*

Donna Lee Fackenthal • *Department of Human Genetics, The University of Chicago, Chicago, IL*

Adam A. Garsa • *Division of Molecular Oncology, Washington University School of Medicine, St. Louis, MO*

Nobumitsu Hanioka • *Laboratory of Health Chemistry, Faculty of Pharmaceutical Sciences, Okayama University, Tokyo, Japan*

Linzy Hill • *SGDP Centre, Institute of Psychiatry, London, UK*

Peter Imle • *Hartwell Center for Bioinformatics and Biotechnology, St. Jude Children's Research Hospital, Memphis, TN*

Federico Innocenti • *Committee on Clinical Pharmacology and Pharmacogenomics, Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL*

Hideto Jinno • *Division of Environmental Chemistry, National Institute of Health Sciences, Tokyo, Japan*

Werner Kalow • *Department of Pharmacology, University of Toronto, Toronto, Canada*

Cristi R. King • *Division of Molecular Oncology, Washington University School of Medicine, St. Louis, MO*

Teri E. Klein • *Department of Genetics, Stanford University School of Medicine, Stanford, CA*

Julian Charles Knight • *Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK*

Pui-Yan Kwok • *Cardiovascular Research Institute, Department of Dermatology, University of California, San Francisco, CA*

Maxwell P. Lee • *Laboratory of Population Genetics, National Cancer Institute, Bethesda, MD*
Sharon Marsh • *Division of Molecular Oncology, Washington University School of Medicine, St. Louis, MO*
Michael Maurer • *Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria*
Howard L. McLeod • *Division of Molecular Oncology, Washington University School of Medicine, St. Louis, MO*
Emma Meaburn • *SGDP Centre, Institute of Psychiatry, London, UK*
Robert Molidor • *Christian Doppler Laboratory for Genomics and Bioinformatics, Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria*
Shogo Ozawa • *Division of Pharmacology, National Institute of Health Sciences, Tokyo, Japan*
Robert Plomin • *SGDP Centre, Institute of Psychiatry, London, UK*
Renato Robledo • *Department of Biology, University of Cagliari, Italy*
Yoshiro Saito • *Division of Biochemistry and Immunochemistry, National Institute of Health Sciences, Tokyo, Japan*
Jun-ichi Sawada • *Division of Biochemistry and Immunochemistry, National Institute of Health Sciences, Tokyo, Japan*
Alexander Sturn • *Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria*
Toshiko Tanaka-Kagawa • *Division of Environmental Chemistry, National Institute of Health Sciences, Tokyo, Japan*
Caroline F. Thorn • *Department of Genetics, Stanford University School of Medicine, Stanford, CA*
Zlatko Trajanoski • *Christian Doppler Laboratory for Genomics and Bioinformatics, Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria*
Dirk van den Boom • *Director of the Molecular Applications Department, Sequenom Inc, San Diego CA*
Matthias Wjst • *Molecular Epidemiology, GSF - Forschungszentrum für Umwelt und Gesundheit, Neuherberg/Munich, Germany*
Ming Xiao • *Cardiovascular Research Institute, University of California, San Francisco, CA*
Hai Yan • *Department of Pathology, Duke University Medical Center, Durham, NC*

I

History and Overview

1

Pharmacogenomics

Historical Perspective and Current Status

Werner Kalow

Summary

Pharmacogenomics is an extension of pharmacogenetics, a science described here in terms of five stages of development: 1) some clinical observations predicted genetic alterations of drug response; 2) additional case discoveries led to the term "pharmacogenetics," a concept broadened by 3) many systemic case studies, and the realization of its wide applicability; 4) came the recognition of systematic pharmacogenetic differences between human populations. Then it became clear that 5) most human drug–repsonse differences were multifactorial, caused by many genetic alterations plus environmental factors. The recognition of these complexities, and the advance of genetics into genomics led to the broader science of pharmacogenomics. This led to plans to create "personalized medicine," that is, making drug use more effective and safer by giving drugs that fit a person's genes. Much of the science of genetics, dealing with gene structure, was changed by the realization that gene expression and thereby gene function was variable; this leads to systematic studies of drug action on genes, reversing the traditional studies of genes affecting drug action.

Finally, the realization that gene–protein variations contribute to most common diseases leads to efforts of creating new drugs that act on these variants.

Key Words: Pharmacogenetics; pharmacogenomics; personalized medicine; populations; gene expression; drug targets.

1. Introduction

Pharmacogenomics is a recent offspring of pharmacogenetics. Both sciences deal with hereditary impacts upon the action of drugs, and their goals are overlapping. Any proper account of the history of pharmacogenomic must include a look at the development of pharmacogenetics. Thus, in this chapter pharmacogenetic history will be outlined first, concentrating on its development in terms of succeeding stages.

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

2. Pharmacogenetics: Its Stages of Development

2.1. Stage 1: Visions and Some Predictive Observations

Some visionaries and some keen observers predated pharmacogenetics as a science. Garrod's 1902 studies ***(1)*** of alcaptonuria and of phenylketonuria indicated to him that there was such a thing as human biochemical individuality. Haldane ***(2)*** summarized his views in 1949 by stating, "It is an advantage to a species to be biochemically diverse. For the biochemically diverse species will contain at least some members capable of resisting any particular pestilence."

There were also some observational forerunners. In 1932, Snyder ***(3)*** described a heritable disability of some people to taste phenylthiocarbamide. In 1943, Savin and Glick ***(4)*** noticed a genetic lack of atropine esterase in some rabbits; these animals died while eating belladonna leaves, whereas most rabbits were not affected. These cases were perceived as isolated observations; they preceded the definition of pharmacogenetics but they helped later investigators to establish pharmacogenetics as a science.

2.2. Stage 2: Pharmacogenetics Lives: Systematic Case Studies

Several separate observations in the 1950s indicated clearly the dependence of drug effects on the genetic constitution of the recipient. The data were convincing because they were based on combinations of biochemical, clinical, and genetic observations. The cases included genetic variation of isoniazid acetylation ***(5)***, failing cholinesterase activity affecting succinylcholine action ***(6)***, and primaquine-caused hemolysis owing to deficiency of glucoses-6-phosphate dehydrogenase ***(7***). All these cases became subjects of subsequent studies that showed that each of the enzymes could vary in many ways because of different mutations.

These clear-cut cases raised in several people the opinion that pharmacological heritability was a clinically important subject. Thus, the American Medical Association (AMA) invited the geneticist Arno Motulsky ***(8)*** to consider the problem; he summarized it in a 1957 paper entitled "Drug Reactions, Enzymes, and Biochemical Genetics." Vogel ***(9)*** in Germany was aware of the same problem and coined the term "Pharmacogenetics." Kalow began to summarize all available knowledge in a book that appeared in 1962 ***(10)***, indicating that the concept of pharmacogenetics had been fundamentally accepted.

2.3. Stage 3: Broadening of Pharmacogenetic Knowledge

In the following years, many centers contributed new data, but all the data represented monogenic variations, i.e., differences between individuals caused by mutations of single genes. Weber's 1997 book ***(11)*** listed 15 variable drug-metabolizing enzymes, 11 variable drug receptors, and 14 other variable proteins in humans that affected drug actions. In 2001, Kalow ***(12)*** counted 42

variable drug-metabolizing enzymes. In short, the knowledge of different kinds of protein variants that may affect drug responses was growing.

Drug-metabolizing enzymes represent the category with the largest number of known variants, which probably has historical and methodological reasons. To measure a change of drug metabolism, all one needs are chemical methods, many of which are dated. To find a drug receptor variation, one has to identify the receptor protein and its gene ***(13)***, processes that require sophisticated procedures of more recent date. Thus, measurements of drug metabolism are older than some other procedures.

In the history of drug metabolism, prominent were the discoveries of genetic variability of the metabolism of debrisoquine ***(14)*** and of sparteine ***(15)***. Subsequent studies indicated that both drugs are metabolized by the same enzyme ***(16)***, which turned out to be the P450 cytochrome CYP2D6 ***(17)***. The enzyme's variations were found to be complex ***(18)***: enzyme activity could be absent because of frameshift mutations, splicing defects, gene deletion, or the presence of a stop codon. The enzyme may function slowly because of various kinds of mutation, whereby some mutations affected only the interaction with specific substrates. Enzyme duplication or multiplication could lead to very fast action.

Many clinical case studies and observations could be mentioned. For example, as summarized by Meyer ***(19)***, patients with deficient CYP2D6 activity experienced exaggerated or prolonged responses to metoprolol, encainide, perhexiline, or thioridazine; however, codeine had no analgesic effect in such cases because it must be activated to morphine by CYP2D6.

More than 60 alleles of CYP2D6 are known today, characterized by different combinations of some 45 mutations ***(20)***. Approximately 60 different drugs are metabolized by CYP2D6 ***(21)***. A Pubmed search indicated that there are more than 2300 publications dealing with CYP2D6. The studies of CYP2D6 helped to give pharmacogenetics the deserved clinical attention.

Genetic failure of drug-metabolizing enzymes can lead to a patient's death. For instance ***(22)***, mercaptopurine or thioguanine have been fatal in cases of failing activity of thiopurine methyltransferase. Regulatory agencies are considering recommendations and official acceptance of pharmacogenetic testing before the administration of dangerous drugs.

2.4. Stage 4: Pharmacogenetic Differences Between Populations

Pharmacogenetics began with the observation of interindividual differences of some drug responses or of drug metabolism that started pharmacogenetics. This recognition, that there are pharmacogenetic differences between populations, truly widened and altered the science. Various older observations that

led to the recognition of a pharmacogenetic difference between human populations were first considered to be odd cases. In 1921, Paskind ***(23)*** injected atropine sulfate into 20 Caucasian and 20 African-American men in Chicago. He found that the drug caused an initial slowing of the heart rate in the white but not the African-American subjects. In 1929, Chen and Poth ***(24)*** measured the pupillary size after applying various mydriatic eye drops into the eyes of a number of people. The increase in size was largest in Caucasians, intermediate in Asians, and smallest in African-Americans; the authors thought that the color of the iris affected its movability.

During World War II, American soldiers stationed in tropical countries received primaquine ***(25)*** as antimalarial prophylaxis; it turned out that only soldiers of African descent developed hemolysis from the administration of primaquine. The explanation came later ***(26)***: The affected soldiers had a genetic deficiency of glucose-6-phosphate dehydrogenase; this deficiency was frequent in Africans because it protected the carrier from malaria, but it was rare in countries without malaria. After discovery of the genetic deficiency of isoniazid acetylation ***(5,27)***, other investigators found substantial interethnic differences in the frequency of this deficiency ***(28)***. The deficiency was rare in Eskimos (Inuit), relatively frequent in Europeans and Africans, and intermediate in East Asians.

In the early 1970s, my laboratory studied the metabolism of amobarbital (an at that time widely used drug) in a class of students ***(29)***. When we did not observe its normal metabolite in 7 of the 140 students, we assumed a laboratory error. When calling the students back for reinvestigation, it turned out that all 7 were of Asian origin and that our first measurements had been correct. At the same time, our laboratory ran tests with debrisoquine ***(30)*** because a genetic variation of its metabolism had just been discovered ***(14)***. Again, we saw a substantial difference of its metabolic destruction between students of Asian and non-Asian origin (the difference was later defined in terms of DNA variation by Swedish investigators ***[31]***). These unexpected observations with amobarbital and debrisoquine caused us to search the literature and to publish in 1982 the first article on interethnic differences in drug metabolism ***(32)***.

In the mean time, studies of interethnic differences in drug response or metabolism have become frequent research projects. Computerized Pubmed lists more than 2000 articles dealing with the combined entries “drug” and “race.” It is now quite clear that the interethnic differences may be divided into two kinds: first, a given mutation of a particular gene may occur with different frequencies in different populations. Second, there are mutations that appear to be specific for a particular population. Some of the differences may be there because they provide a population with a biological advantage; however,

some mutations may simply differ because they have arisen in a population after it separated from others.

This difference is suggested by some overview data ***(12)***. The occurrence of 11 mutations that affect the function of the P450 cytochrome CYP2D6 was tested by various investigators in populations from Europe, China, Japan, and Africa. Of these 11 mutations, Europeans carried 7, Chinese 4, Japanese 3, and Africans 2. Only one mutation (G4268C) was found in all countries, suggesting that it arose before humanity separated into different ethnicities.

Inter-ethnic differences occur frequently. As stated previously ***(12)***, 42 drug-metabolizing enzymes have shown pharmacogenetic variability within one or other population. When checking the literature for the occurrence of interethnic differences between these variants, researchers discovered that 28 (66%) showed such differences. This percentage is considered high, particularly when noting the fact that interethnic comparisons had never been made for many drugs. In short, if we view a pharmacogenetic variation between people, it is likely absent in other populations or occurring with a different frequency. Is this a rule that holds for all mutations in any gene?

2.5. Stage 5: The Rise of Multifactorial Pharmacogenetics

Differences between people in their response to drugs are regular occurrences. This observation was formalized in 1927 by introduction of the term and the concept of ED50 ***(33)***; it indicates the dose of a drug sufficient to produce a given effect in 50% of the members of a population. In other words, all drug effects are variable. This result is not surprising because there are numerous factors that can affect a drug response.

As a simple example, let us consider the rate of metabolism of any particular drug. The metabolism may fail because of a genetic change of the enzyme structure. Perhaps the metabolism failed because not enough enzyme was formed, perhaps because of low gene expression or because of a failure of transcription or translation. Was there the absence of an inducing or regulating hormone, or was the enzyme degraded too quickly? Perhaps a genetic abnormality of the promoting region prevented the normal response to the inducer. Perhaps the drug could not reach the enzyme because it was bound somewhere else or a transporter was missing. Thus, even a single step in the drug's fate may be complex and affected by many genes; the genes may interact, and environmental factors also may contribute to the variation.

The causes of most differences generally remain uninvestigated, but the presence of both genetic and environmental causes is common ***(34)***. It is of considerable interest to know the relative contribution of the two causes. The classical method of investigation, used prominently, for example, by Vesell,

consisted of twin studies *(35)*: that is, the magnitude of differences between the two members of a pair of identical and a pair of fraternal was measured. Repeating such studies in many twins and averaging and comparing the differences allowed a calculation of heritability. Unfortunately, the recruiting of a sufficient number of twins often is difficult.

However, a simpler method is now available *(36)* that is based on the fact that one can give a drug repeatedly to a person and measure each response and the difference between the responses. When giving a drug at an appropriate interval two or more times to a group of people, one can measure two magnitudes: first, one can calculate the average and standard deviation of the response differences between the first and subsequent applications in the same people; second, one can equally calculate and record the difference between subjects. Let us designate the standard deviations of the within-subject variations as SD_w, of the between-subjects as SD_b, and the genetic component of the between-subject variation as r_{GC}. Squaring the standard deviations, the genetic component is then calculated by the following equation:

$$r_{GC} = (SD_b^2 - SD_w^2)/SD_b^2 \textit{ [37]}).$$

A value close to 1.0 indicates overwhelming heredity, close to 0 indicates mostly environmental influence. A recent example of the use of this method has been the assessment of genetic and environmental determinants of cytochrome CYP3A4 activity *(38)*.

Often forgotten is the fact that there may be a clinically significant drug response difference between two populations even if the average response differences are small, perhaps not even statistically significant *(39)*. If population data are represented by a normal distribution (Gauss) curve, the persons with abnormal responses may be represented by one of the edges of the curve.

For example, let us consider a normally distributed metabolic destruction rates of drug X. Assume that 2% of the persons have a destruction rate low enough to suffer toxicity from the drug. In another group of people, destruction of the same drug may have a somewhat lower average rate, a fact that is immaterial for most subjects; however, if the distribution curves have equal spread in the two populations, many more people in the second population will have the critically low drug destruction rate and will be intoxicated than in the first population (**Fig. 1**). In short, the difference between the edges of the distribution curves may be of clinical and statistical significance even if the averages are similar.

3. Pharmacogenomics

As described previously, pharmacogenetics began with the study of single gene differences between individuals but developed into a broad science. Methodological advances expanded the science further into pharmacogenomics; one

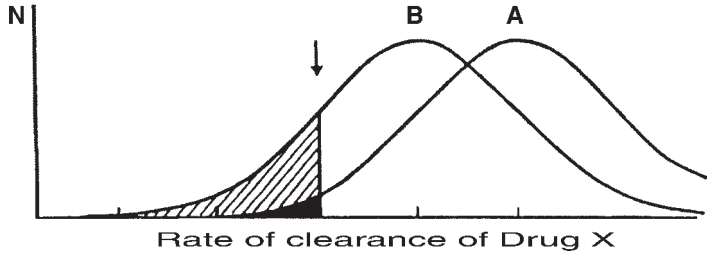


Fig. 1. Variation of drug clearance in two populations. The figure illustrates a hypothetical case showing a 20-fold variation of drug clearance between the individuals within each population, and a twofold difference between the population means. Let us assume that a slow clearance will tend to produce overdose toxicity in affected subjects. Obviously, the proportion of such subjects would be much larger in population B than in A.

may say that pharmacologists followed the geneticist's adoption of genomic techniques. The consequence should be a better understanding of the multiplicities and complexities of drug–gene interactions; not only may genes affect drug action, but drugs may affect gene function. Pharmacogenomics represents attempts by researchers to medically use these new understandings together with the old ones. The hope is to optimize the efficacy of drugs, to minimize adverse drug reactions, and to facilitate drug discovery, development, and approval.

4. The Aims of Pharmacogenomics

4.1. Aim 1: Creation of a Basis for Personalized Medicine

Current medicine is based on statistical likelihood and often fails the individual. The incidence of serious or fatal drug reactions depends in many or most cases on genetic variation. Studies from U.S. hospitals ***(40)*** suggested that 6.7% of patients had serious adverse and 0.32% had fatal, drug reactions. The latter caused approx 100,000 deaths per year in the United States. Many researchers hope that adverse reactions or therapeutic failures will be eliminated by the introduction of personalized medicine ***(41)***, meaning that the drug to be given to a patient will be determined by the patient's genes.

To reach this aim, we must learn much more about the genetic variants that may affect drug action. The most frequently occurring genetic variants are single-nucleotide polymorphisms (SNPs ***[42]***). The human genome contains approx 3 billion basepairs, and SNPs occur on the average in approx 1 per 1000 bases; thus, they cause genetic variation of many human proteins. These variants are important objects of study because human individuals usually differ from each other by less than 1% of their genes; thus, SNPs are important.

Besides being most common, SNPs are the most technically accessible class of genetic variants. Using genomic methods, high-density maps of SNPs can be created and hopefully used as distinguishing markers of xenobiotic response, even if the drug target is not specifically identified.

By correlating SNPs and drug response data, one will have gained an ability to predict drug efficacy or toxicity within reasonable limits for any individual ***(43)***. Although variants other than SNPs are able to affect drug responses, this detection will happen less frequently than the SNP-related ones. Nevertheless, screening of the genome for all known and for unknown pharmacogenomics-related variations will be tasks for years to come.

Personalized medicine can also be said to be based on the identification of biomarkers ***(44)***. They extend to a broad variety of indirect manifestations of underlying, but often unrecognized, sequence variations. Because of their variety, their study or discovery may require various kinds of investigations. Many published examples concern cancers. To name a few, prostate cancer is subject to transcriptional regulation ***(45)***, and there are prostate-specific antigens ***(46)***. Cervical neoplasia may depend on the human papilloma virus ***(47)***. Many other examples apply to cardiovascular diseases. For example, high levels of C-reactive protein or interleukin 6 affect ischemic heart disease and its mortality risk ***(48)***. Low-density lipoprotein cholesterol is used to profile cardiac risks ***(49)***. A strange example in a different field is the fact that excessive alcohol consumption can be assessed by the serum level of carbohydrate-deficient transferrin ***(50)***. Thus, the complexity of biomarkers is frightening; even if we have reached personalized medicine, it will not be without problems.

4.2. Aim 2: Drugs Affecting Gene Expression

For a gene to form a protein, its DNA has to be converted into RNA, which acts within a ribosome. Various studies have shown that the amount of RNA can vary, indicating functional variation of the gene, quantified in terms of gene expression ***(51)***. Microarray experiments represent a genomic technique that has yielded expression information of thousands of genes.

The understanding that gene expression is variable has changed genetics, which was for a long time concerned only with structural differences between genes. We now know that gene interaction may mean that one gene affects expression and function of other genes.

Furthermore, gene expression may be changed by hormones, by disease, by food, or by drugs. In pharmacology, a good example of variable gene expression is drug-increased drug metabolism ***(52)***, a newer one is drug addiction caused by drugs that act on genes in brain ***(53)***. We sometimes do not know whether a drug action is the result of the drug affecting a protein or a gene.

This question may be answered by using gene expression changes. Using repeat studies, one can measure expressions obtained before and after exposure to a drug. Any difference of expression tells which, if any, genes are affected by the drug ***(54)***. One can test gene expression in specific cells, be it leukocytes or liver cells. If tested in brain cells, one may learn whether the drug affects a gene of interest in brain.

4.3. Aim 3: Identification of New Targets for Future Drugs

Common diseases are usually caused by a combined action of several or of many genes, in addition to environmental influences ***(55)***. Genomic methods that allow an investigation of numerous genes may help to identify a gene that contributes importantly to a disease. If so, one may look for a chemical that targets that gene or its protein product; this chemical may then become a drug that helps to combat the disease. This kind of effort has been called a search for "drugable targets," that is, for targets that can be approached by small molecules of the size of drugs. Thus, genomic studies may lead to new medical therapies.

Because the set of genes responsible for a given disease may differ somewhat between people, the targets able to be drugged may not be exactly the same in different subjects ***(56)***. This hypothesis is not classical pharmacogenetics, but it explains why a given drug may cure the disease only in some people. Genomic studies that tell ahead of time which person may benefit from the drug are a part of personalized medicine.

The search for disease-causing genes may also be helped by gene expression studies ***(57)***. By comparing genes in the presence and in the absence of a disease, one may get some indication of which genes are affected by the disease, or which genes cause the disease. In exceptional cases, one may be able to compare gene expression in a person before and after onset of the disease. Usually, it is necessary to compare groups of healthy and diseased subjects because there will be differences of gene expression that are unrelated to the disease.

Another set of target studies is the attempt to create new antibiotics or infection-fighting drugs by investigating the genetic structure of bacteria or other infectious agents. Thereby, one may identify potential targets for new drugs. Even if the targets are known, finding such new drugs may require intensive systematic searches.

5. Conclusion

Genetics and pharmacology are two sciences that interact in many different ways. The study of such interactions was aroused by some simple observations that indicated that monogenic differences could cause persons to respond dif-

ferently to a drug. Unfortunately, both the effects of drugs and the effects of genes can vary a great deal, and the interactions of these two turned out to be often so complex that they are frequently hard to understand. Nevertheless, pharmacogenomics is the science that studies these interactions. Its purpose is to unlock some of the difficulties, to use as many facts as possible to improve medicine, and thereby to help all human beings.

References

1. Garrod, A. E. (1931) *Inborn Factors in Disease: An Essay*. Oxford University Press, New York, NY.
2. Haldane, J. B. S. (1932) *The Causes of Evolution*. Princeton University Press, London (Reprinted by Longmans, Green 1990).
3. Snyder, L. H. (1932) Studies in human inheritance. IX. The inheritance of taste sensitivity in man. *Ohio J. Sci.* **32,** 436–440.
4. Savin, P. B. and Glick, D. (1943) Hydrolysis of atropine by esterase present in rabbit serum. *Proc. Natl. Acad. Sci. USA* **29,** 55.
5. Hughes, H. B., Biehl, J. P., Jones, A. P., and Schmidt, H. L. (1954) Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. *Am. Rev. Tuberculosis* **70,** 266–273.
6. Kalow, W. (1956) Familial incidence of low pseudocholinesterase level. *Lancet* **2,** 576–577.
7. Beutler, E. (1959) The hemolytic effect of primaquine and related compounds: a review. *Blood* **24,** 103–139.
8. Motulsky, A. G. (1957) Drug reactions, enzymes, and biochemical genetics. *JAMA* **165,** 835–837.
9. Vogel, F. (1959) Moderne Probleme der Humangenetic. *Ergebnisse der Inneren Medizin und Kinderheilkunde* **12,** 65–126.
10. Kalow, W. (1962) *Pharmacogenetics. Heredity and the Response to Drugs*. W.B. Saunders Company, London, UK.
11. Weber, W. (1997) *Pharmacogenetics*. Oxford University Press, New York, NY.
12. Kalow, W. (2001) Interethnic differences in drug response, In: Pharmacogenomics (Kalow, W., Meyer, U. A., and Tyndale R., eds.), Marcel Dekker Inc., New York, NY, pp. 109–134.
13. Satoh, M. and Minami, M. (1995) Molecular pharmacology of the opioid receptors. *Pharmacol. Ther.* **68,** 343–365.
14. Mahgoup, A., Dring, L. G., Idle, J. R., Lancaster, R., and Smith, R. L. (1977) Polymorphic hydroxylation of debrisoquine in man. *Lancet* **2,** 584–586.
15. Eichelbaum, M., Spanbrucker, N., Steinke, B., and Dengler, H. J. (1979) Defective N-oxidation of sparteine in man: a new pharmacogenetic defect. *Eur. J. Clin. Pharmacol.* **16,** 183–187.
16. Otton, S. V., Inaba, T., Mahon, W. A., and Kalow, W. (1982), In vitro metabolism of sparteine by human liver: Competitive inhibition by debrisoquine. *Can. J. Physiol. Pharmacol.* **60,** 102–105.

17. Eichelbaum, M., Bertilsson, L., Sawe, B. J., and Zekorn, C. (1982). Polymorphic ocidation of sparteine and debrisoquine: related pharmacogenetic entities. *Clin. Pharmacol. Ther.* **31,** 184–186.
18. Marez, D., Legrand, M., Sabbagh, N., Lo Guidice, J. M., Spire, C., Lafitte, J. J., et al (1997) Polymorphism of the cytochrome P450 CYP2D6 gene in a European population: characterization of 48 mutations and 53 alleles, their frequencies and evolution. *Pharmacogenetics* **7,** 197–202.
19. Meyer, U. A. (2001) Pharmacogenetics: clinical viewpoints, in *Pharmacogenomics* (Kalow, W., Meyer, U. A., and Tyndale, R., eds.) Marcel Dekker Inc., New York, NY, pp. 135–150.
20. Daly, A. K., Brockmoller, J., Broly, F., et al. (1996) Nomenclature for human CYP2D6 alleles. *Pharmacogenetics* **6,** 193–201.
21. Kalow, W. and Grant, D. M. (2001). *Pharmacogenetics*, in The metabolic and molecular bases of inherited disease (Scrivener, C. R., Beaudet, A. L., Sly, W. S., et al., eds.), McGraw-Hill, New York, NY, pp. 225–255.
22. Weinshilboum, R. M. (2001) Thiopurine pharmacogenetics: clinical and molecular studies of thiopurine methyltransferase. *Drug Met. Dispos.* **29,** 601–605.
23. Paskind, H. A. (1921) Some differences in response to atropine in white and coloured races. *J. Lab. Clin. Med.* **7,** 104–108.
24. Chen, K. K. and Poth, E. J. (1929) Racial differences as illustrated by the mydriatic action of cocaine, enphthalmine and ephedrine. *J. Pharmacol. Exp. Ther.* **36,** 429–434.
25. Beutler, E. (1993) Study of glucose-6-phosphate dehydrogenase: history and molecular biology. *Am. J. Hematol.* **42,** 53–58.
26. Motulsky, A. G., (1960), Metabolic polymorphisms and the role of infectious diseases in human evolution. *Human Biol.* **32,** 28–32.
27. Hughes, H. B., Biehl, J. P., Jones, A. P., and Schmidt, L. H. (1954) Metabolism of isoniazid metabolism in man as related to the occurrence of peripheral neuritis. *Am. Rev. Tuberculosis* **70,** 266–273.
28. Sunahara, S, Urano, M., and Ogawa, M. (1961) Genetical and geographical studies on isoniazid inactivation. *Science* **134,** 1530–1531.
29. Kalow, W., Tang, B-K., Kadar, D., Endrenyi, L., and Chan, F. Y. (1979) A method to study drug metabolism in populations: racial differences in amobarbital metabolism. *Clin. Pharmacol. Ther.* **6,** 766–776.
30. Kalow, W., Otton S.W., Kadar, D., Endrenyi, L., and Inaba, T. (1980) Ethnic difference in drug metabolism: debrisoquine 4-hydroxylation in caucasians and orientals. *Can. J. Physiol. Pharmacol.* **58,** 1143–1144.
31. Bertilsson, L., Lou, Y. Q., Du, Y. L., et al. (1992) Pronounced differences between native Chinese and Swedish populations in the polymorphic hydroxylation of debrisoquine and S-mephenytoin. *Clin. Pharmacol. Ther.* **51,** 388–397.
32. Kalow, W. (1982) Ethnic differences in drug metabolism. *Clin. Pharmacokinetic.* **7,** 373–400.
33. Trevan, J. W. (1927) The error of determination of toxicity. *Royal Soc. London Proc. 1, Ser. B* **101,** 483–514.

34. Motulsky, A. G. (1978) Multifactorial inheritance and heritability in pharmacogenetics. *Human Genet.* **44,** (Suppl 1), 7–11.
35. Vesell, E. S. (1974) Polygenic factors controlling drug response. *Med. Clin. North Am.* **58,** 951–963.
36. Kalow, W., Endrenyi, L., and Tang, B. K. (1999) Repeat administration of drugs as a means to assess the genetic component in pharmacological variability. *Pharmacology*. **58,** 281–284.
37. Kalow, W., Ozdemir, V., Tang, B. K., Tothfalusi, L., and Endrenyi, L. (1999) The science of pharmacological variability: an essay. *Clin. Pharmacol. Ther.* **66,** 445–447.
38. Ozdemir, V., Kalow, W., Tang, B. K, et al. (2000) Evaluationof the genetic component of variability of CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics* **10,** 373–388.
39. Kalow, W. and Bertilsson, L. (1994) Interethnic factors affecting drug response. *Adv. Drug Res.* **25,** 1–59.
40. Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279,** 1200–1205.
41. Kalow, W. (2002) Pharmacogenetics and personalised medicine. Fund. *Clin. Pharmacol*. **16,** 337–342.
42. Pennisi, E. (1998) Using the wildly popular genome markers called SNPs to track genes may be less straightforward than researchers expected. A closer look at SNPs suggests difficulties. *Science* **281,** 1787–1789.
43. Pfost, D. R., Boyce-Jacino, M. T., and Grant, D. M. (2000) A SNP snapshot: pharmacogenetics and the future of drug therapy. *Trends Biotechnol.* **18,** 334–338.
44. Silber, B. M. (2001) Pharmacogenomics, biomarkers, and the promise of personalized medicine, in *Pharmacogenomics* (Kalow, W., Meyer, U. A., Tyndale, R., eds.) Marcel Dekker Inc., New York, NY.
45. Schalken, J. A., Hessels, D., and Verhaegh, G. (2003) New targets for therapy in prostate cancer: differential display code 3 (DD3(PCA3)), a highly prostate cancer-specific gene. *Urology* **62,** 34–43.
46. Lilja, H. (2003) Biology of prostate-specific antigen. *Urology* **62,** 27–33.
47. Wolf, J. K., Franco, E. L., Arbeit, J. M., et al. (2003) Innovations in understanding the biology of cervical cancer. *Cancer* **98,** 2064–2069.
48. Schwartz, R. S., Bayer-Genis, A., Lesser, J. R., Sangiorgi M., and Conover, C. A. (2003) Detecting vulnerable plaque using peripheral blood: inflammatory and cellular markers. *J. Interv. Cardiol.* **16,** 231–242.
49. Dornbrook-Lavender, K. A. and Pieper, J. A. (2003) Genetic polymorphisms in emerging cardiovascular risk factors and response to statin therapy. *Cardiovasc. Drugs Ther*. **17,** 75–82.
50. Montalto, N. J.and Bean, P., (2003) Use of contemporary biomarkers in the detection of chronic alcohol use. *Med. Sci. Monit*. **9,** RA285–RA290.
51. Burgess, J. K. (2001) Gene expression studies using microarrays. *Clin. Exp. Pharmacol. Physiol.* **28,** 321–328.

52. Rae, J. M., Johnson, M. D., Lippman, M. E., and Flockhart, D. A. (2001) Rifampin is a selective, pleiotropic inducer of drug metabolism genes in human hepatocytes: studies with cDNA and oligonucleotide expression arrays. *J. Pharmacol. Exp. Ther.* **299,** 849–857.
53. Rahman, S. and Miles, M. F. (2001) Identification of novel ethanol-sensitive genes by expression profiling. *Pharmacol. Ther.* **92,** 123–134.
54. Steiner, S. and Anderson, N. L. (2000) Expression profiling in toxicology—potentials and limitations. *Toxicol. Lett.* **112–113,** 467–471.
55. King, R. A., Rotter, J. I., and Motulsky, A. G. (eds.) (2002) *The Genetic Basis of Common Diseases* (2nd ed.). Oxford University Press, New York, NY.
56. Thomson, G. (2001) Mapping of disease loci, in *Pharmacogenomics* (Kalow, W., Meyer U.A., Tyndale, R., eds.), Marcel Dekker Inc., New York, NY.
57. Bals, R. and Jany, B. (2001) Identification of disease genes by expression profiling. *Eur. Respir. J.* **18,** 882–889.

II

Functional Analysis of Gene Variation

2

Transfection Assays With Allele-Specific Constructs

Functional Analysis of UDP–Glucuronosyltransferase Variants

Hideto Jinno, Nobumitsu Hanioka, Toshiko Tanaka-Kagawa, Yoshiro Saito, Shogo Ozawa, and Jun-ichi Sawada

Summary

Adverse drug reactions (ADRs) are a major clinical problem. A rapidly growing body of evidence suggests that genetic factors, at least in part, determine individual susceptibility to ADRs. A large number of pharmacogenetic studies have identified a number of polymorphisms as predictors of drug efficacy and/or adverse events. These candidate markers should be investigated further to ascertain the underlying mechanism of action, for example, changes in the kinetic parameters of an enzyme, or transcriptional activity of a promoter region. In this chapter, we describe a transient transfection assay for the functional characterization of naturally occurring variants of UDP–glucuronosyltransferase (UGT) 1A1. This phase II drug metabolizing enzyme is involved in the glucuronidation of SN-38, an active metabolite of the anti-cancer drug irinotecan. Single-nucleotide polymorphisms of the *UGT1A1* gene have been correlated to irinotecan-induced ADRs. Variant UGT1A1s are heterologously expressed in COS-1 cells and characterized in terms of the level of protein expression and enzyme kinetics.

Key Words: UDP–glucuronosyltransferase 1A1; single-nucleotide polymorphism; adverse drug reactions; SN-38 glucuronidation; kinetic analysis.

1. Introduction

UDP–glucuronosyltransferase (UGT) 1A1 catalyzes the glucuronidation of bilirubin, thereby rendering it soluble for excretion. A genetic defect in *UGT1A1*, therefore, can result in a phenotype of unconjugated hyperbilirubinemia, Crigler–Najjar syndrome, and Gilbert's syndrome. UGT1A1, along with UGT1A7 and UGT1A9, also is known to play a dominant role in the glucuronidation of SN-38, an active metabolite of the anti-cancer drug irinotecan ***(1,2)***. Pharmacogenetic studies have revealed that several polymor-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

phisms in *UGT1A1* affect the pharmacokinetics of irinotecan/SN-38 ***(3)*** and consequently the incidence of adverse side effects of SN-38, such as severe diarrhea and neutropenia ***(4,5)***. Large ethnic differences exist in *UGT1A1* polymorphisms, and non-synonymous variations in the coding region have been found in Japanese/Asian populations at relatively high frequencies; 211G>A (amino acid substitution of G71R), 247T>C (F83L), 686C>A (P229Q), and 1456T>G (Y486D) ***(6,7)***. We have characterized the functional alterations for some of these UGT1A1 variants using the heterologously expressed recombinant proteins ***(8)***.

A simple way of evaluating the affect of a single nucleotide polymorphism on enzyme function is to perform transient transfection assays in COS-1 cells. Western blot analysis, with anti-UGT1A antibody, is used to determine protein expression levels of the variants, which often correlate with protein stability. Enzyme kinetic analysis is used to investigate the functional impact of amino acid substitutions.

2. Materials

2.1. Plasmid Construction

2.1.1. TA Cloning

1. Human adult normal liver complementary deoxyribonucleic acid (cDNA; BioChain Institute Inc., Hayward, CA).
2. TaKaRa LA Taq DNA polymerase (Takara, Kyoto, Japan).
3. TA cloning kit (Invitrogen, Carlsbad, CA).
4. Restriction enzymes: *Not*I and *Bam*HI (Takara).
5. Calf intestinal alkaline phosphatase (Takara).
6. DNA ligation kit ver.2 (Takara).
7. pcDNA 3.1 (-) vector (Invitrogen).
8. Library efficiency *Escherichia coli* DH5α-competent cells (Invitrogen).
9. *E. coli* culture media: luria broth (LB) medium and LB agar plate with 50 μg/mL kanamycin or 100 μg/mL ampicillin. Miller's LB powder (Invitrogen) or LB agar powder (Invitrogen) is dissolved in distilled water and autoclaved. The media is cooled to approx 50°C before adding the antibiotic from a 1000X stock solution (kanamycin, 50 mg/mL or ampicillin, 100 mg/mL).
10. ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction kits version 2.0 (Applied Biosystems, Foster City, CA).

2.1.2. Site-Directed Mutagenesis

1. QuikChange site-directed mutagenesis kit (Stratagene, La Jolla, CA).
2. Mutagenic oligonucleotide primers (**Table 1**) were obtained from Proligo Japan, Kyoto, Japan.

Table 1
Primers Used for Plasmid Construction

Purpose	Primer name	Sequence[a]
TA cloning	UGT1A1_F	5'-CAAAGGCGCCATGGCTGT-3'
	UGT1A1_R	5'-CTTATTTCCCACCCACTTCTCA-3'
Site-directed mutagenesis	Mut_G71R_F	5'-CCTCGTTGTACATCAGAGACA**A**GAGCATTTTACA CCTTGAAG-3'
	Mut_G71R_R	5'-CTTCAAGGTGTAAAATGCTC**T**GTCTCTGATGTA CAACGAGG-3'
	Mut_F83L_F	5'-CGTACCCTGTGCCA**C**TCCAAAGGGAGGATGTG-3'
	Mut_F83L_R	5'-CACATCCTCCCTTTGGA**G**TGGCACAGGGTACG-3'
	Mut_P229Q_F	5'GCGACGTGGTTTATTCCC**A**GTATGCAACCCT TGCCTC-3'
	Mut_P229Q_R	5'-GAGGCAAGGGTTGCATAC**T**GGGAATAAACC ACGTCGC-3'
	Mut_Y486D_F	5'CCTCACCTGGTACCAG**G**ACCATTCCTTGGACG-3'
	Mut_Y486D_R	5'-CGTCCAAGGAATGGT**C**CTGGTACCAGGTGAGG-3'

[a]Bold letters show the nucleotides exchanged.

2.2. Transient Expression of UGT1A1s in COS-1 Cells

1. COS-1 cells from the Health Science Research Resources Bank (Osaka, Japan).
2. Dulbecco's Modified Eagle's Medium (DMEM; Invitrogen) supplemented with 10% fetal bovine serum (FBS; Invitrogen).
3. Solution of trypsin (0.25%) and ethylenediamine tetraacetic acid (1 m*M*) from Invitrogen.
4. Opti-MEM (Invitrogen).
5. Lipofectamine 2000 reagent (Invitrogen).
6. Phosphate-buffered saline.
7. Buffered sucrose: 0.25 *M* sucrose, 5 m*M* *N*-hydroxyethylpiperazine-*N*'-2-ethanesulfonate, pH 7.4.

2.3. SDS-PAGE and Western Blotting

1. 10% Polyacrylamide gel (PAGE; READYGELS J) from Bio-Rad Laboratories, Inc. (Hercules, CA).
2. Running buffer (10X): 250 m*M* Tris, 1.92 *M* glycine, and 1.0% (w/v) sodium dodecyl sulfate (SDS). Store at room temperature.
3. Sample buffer (2X; Wako Pure Chemical Industries, Ltd. Osaka, Japan): 0.125 *M* Tris-HCl, 4% (w/v) SDS, 20% (w/v) glycerol, 0.002% (w/v) bromophenol blue, 10% (w/v) 2-mercaptoethanol. Store at 4°C.

4. Molecular weight markers: MagicMark XP Western protein standards (Invitrogen).
5. Polyvinylidene diflouride (PVDF) membrane (ATTO Corp., Tokyo, Japan).
6. Blotting buffer: 0.1 *M* Tris, 0.192 *M* glycine, 5% methanol.
7. Tris-buffered saline with Tween-20 (TBS-T): prepare 10X stock with 1.37 *M* NaCl, 0.2 *M* Tris-HCl, pH 7.6, 1% Tween-20.
8. Blocking buffer: 5% (w/v) non-fat dried milk (skim milk: Difco, BD Bioscience, Franklin Lakes, NJ) in TBS-T.
9. Primary antibody: rabbit anti-human UGT1A (BD Gentest, Woburn, MA), rabbit anti-calnexin polyclonal antibody (Stressgen Biotechnologies Inc., San Diego, CA).
10. Secondary antibody: donkey anti-rabbit Ig coupled to horseradish peroxidase (Amersham Biosciences, Piscataway, NJ).
11. Enhanced chemiluminescent (ECL) plus reagents from Amersham Biosciences.
12. Stripping buffer: 2% (w/v) SDS, 0.1 *M* 2-mercaptoethanol, 62.5 m*M* Tris-HCl, pH 6.7.

2.4. Assay for SN-38 Glucuronidation

1. Reaction mixture for enzyme assay: 500 m*M* Tris-HCl buffer, pH 7.4, 100 m*M* $MgCl_2$. SN-38 (kindly provided from Yakult Honsha Co. Ltd. Tokyo, Japan) is dissolved in dimethyl sulfoxide/0.05 *N* NaOH (50:50) at 2.5–150 µ*M*. UDP–glucuronic acid (Wako) is dissolved in distilled water at 50 m*M*.
2. Standards: a stock solution (2.5 m*M*) of SN-38 glucuronide (kindly supplied by Yakult Honsha Co. Ltd.) is dissolved in 5 mL of methanol. A working solution (0.5–250 n*M*) for calibration curves is prepared by the serial dilution of the 2.5 m*M* stock solution with high-performance liquid chromatography (HPLC) elution buffer.
3. Termination solution for enzyme reaction: 10% (w/v) $HClO_4$ is prepared by dilution of the 60% (w/v) $HClO_4$ (Wako).
4. HPLC mobile phase: 50 m*M* KH_2PO_4 containing 3 m*M* sodium 1-octanesulfonate, pH 2.5.
5. Acetonitrile (HPLC grade, Wako).
6. Methanol (HPLC grade, Wako).

3. Methods

3.1. Plasmid Construction

We describe here the cDNA cloning of UGT1A1 by the TA cloning method because it is one of the most well-known polymerase chain reaction (PCR)-based cDNA cloning methods. As an alternative strategy, we have also successfully applied the Gateway recombinational cloning method (Invitrogen) for the functional characterization of UGT1A9 and UGT1A10 variants ***(9,10)***.

3.1.1. TA Cloning

1. UGT1A1 cDNA is amplified by PCR from human liver cDNA. The 100-µL amplification mixture contains 5 U of TaKaRa LA Taq DNA polymerase, 1X LA

PCR Buffer II, 1.5 m*M* $MgCl_2$, 50 µ*M* dNTP, and 0.2 µ*M* each of forward and reverse primers (**Table 1**; *see* **Note 1**). The cycling parameters are as follows: initial denaturation at 95°C for 1 min, followed by 30 cycles of denaturation at 95°C for 30 s and annealing/extension at 67°C for 2 min. Finally, the reaction is terminated by a 10-min extension at 72°C.

2. The PCR product is cloned into pCR 2.1 vector using a TA cloning kit. The 10-µL ligation reaction consists of 1 µL of 10X ligation buffer, 2 µL of the PCR product, 50 ng of pCR 2.1 vector, and 4 U of T4 DNA ligase. The reaction mixture is incubated overnight at 14ºC. Chemically competent *E. coli* TOP10F' cells are then transformed with the ligation mixture and plated on LB medium containing 50 µg/mL kanamycin. Ten colonies are picked at random and plasmid DNA prepared from each. The insert DNA is then sequenced on both strands using an Applied Biosystems 3700 sequencer employing the BigDye Terminator Cycle Sequencing Ready Reaction kit, version 2.0.
3. The resulting plasmid containing the correct insert (designated as pCR-UGT1A1/WT) is double digested with *Not*I and *Bam*HI for 4 h at 37ºC in 0.5X Takara universal buffer K containing 0.01% BSA. Subsequently, the UGT1A1 cDNA fragment is ligated into a mammalian expression plasmid pcDNA3.1(-), which is previously digested with the same enzymes followed by the treatment with alkaline phosphatase, calf intestine for 30 min at 37ºC (*see* **Note 2**). TE buffer (10 µL) containing 300 ng of the UGT1A1 fragment and 100 ng of linearized plasmid DNA is mixed with 10 µL of enzyme solution of DNA ligation kit, version 2. The reaction mixture is incubated for 30 min at 16ºC and then chemically competent *E. coli* DH5α cells are transformed with the ligation mixture. The cells are then plated on LB medium containing 100 µg/mL ampicillin.

3.1.2. Site-Directed Mutagenesis

1. Mutagenic primers (**Table 1**) are designed using the following criteria: 1) the melting temperature (T_m) of the primers (25–45 bp) is ≥78°C, 2) GC content of the primer is ≥40%, 3) the primer terminates in one or more C or G bases, 4) the desired mutation is in the middle of the primer with 10 to15 bases of correct sequence on both sides. T_m is calculated here as follows: 81.5 + 0.41 × GC content (%) – 675/primer length (bp) – % mismatch.
2. Mutations are introduced using a PCR-based site-directed mutagenesis kit (QuikChange site-directed mutagenesis kit). The reaction mixture (50 µL) consist of 5 µL of 10X reaction buffer, 1 µL of dNTP mix, 10 ng of pCR-UGT1A1/WT, 125 ng each of the forward and reverse mutagenic primers, and 2.5 U of *PfuTurbo* DNA polymerase (*see* **Note 3**). The cycling parameters are as follows: denaturation at 95ºC for 30 s, followed by 12 cycles of 95ºC for 30 s, 55ºC for 1 min, and 68ºC for 6 min. The extension time corresponds to 1 min per kb of DNA template.
3. The methylated parental plasmid DNA is digested with *Dpn* I for 1 h at 37ºC. XL1-Blue supercompetent cells are transformed with 1 µL of the reaction mixture and plated on LB medium containing 50 µg/mL kanamycin.
4. Five colonies for each UGT1A1 variant are picked at random and plasmid DNA prepared. The insert DNA is then sequenced on both strands. DNA verified as correct is then subcloned into pcDNA3.1(-) as described previously.

3.2. Transient Expression of UGT1A1s in COS-1 Cells

1. COS-1 cells are maintained in DMEM medium supplemented with 10% FBS and split at a ratio of 1:5 to 1:10 upon reaching 80 to 90% confluence. The day before transfection, the cells were plated in 100-mm culture dishes at a density of 5.5×10^4 cells/cm^2, or 3.0×10^6 cells/dish. For each UGT1A1 variant, three 100-mm dishes are required for the Western blot analysis and the enzyme assay.
2. Just before transfection, the culture medium is replaced with 8 mL of prewarmed Opti-MEM (*see* **Note 4**). The diluted plasmid DNA (14 µg in 810 µL of Opti-MEM) and the diluted Lipofectamine 2000 reagent (48 µL in 810 µL of Opti-MEM) are combined and incubated for 20 min at room temperature. The resulting plasmid DNA–Lipofectamine 2000 complex is then added directly to each dish. After 4 h, the medium is replaced with DMEM medium supplemented with 10% FBS.
3. Forty-eight hours after transfection, the cells are washed twice with 5 mL of ice-cold phosphate-buffered saline and harvested in 2 mL of buffered sucrose using a cell scraper. The cells are transferred into a 15-mL conical tube, precipitated by centrifugation at 1500*g* for 10 min at 4ºC, and resuspended in 1 mL of buffered sucrose.
4. The chilled cell suspension is sonicated for 1 min at a pulse cycle of 1 s, using an ultrasonic processor VC130 equipped with a 3-mm probe (Sonics, Newtown, CT), followed by centrifugation at 105,000*g* for 60 min at 4ºC.
5. The resulting membrane fractions are resuspended in appropriate volume of buffered sucrose; addition of 100 µL of buffered sucrose per 100-mm dish will routinely produce approx 15 mg of protein per milliliter of suspension. The protein concentration of each membrane fraction is adjusted to 10 mg of protein per milliliter with buffered sucrose. Membrane fractions are stored at –80ºC until used for Western blotting and SN-38 glucuronidation assay.

3.3. SDS-PAGE and Western Blotting

1. This protocol is intended to use the Bio-Rad Mini PROTEAN 3 cell and Trans-Blot SD Semi-Dry Electrophoretic Transfer Cell. Any other apparatus would work well under similar conditions.
2. SDS-PAGE samples are prepared by adding equal amounts of 2X sample buffer to the membrane fractions (10 mg protein/mL), boiling for 5 min and cooling to room temperature.
3. A precast 10% gel (Ready Gels from Bio-Rad) is assembled in a Mini PROTEAN 3 electrophoresis module, and the 1X running buffer is added to the inner chamber and the mini tank. Each well is rinsed thoroughly with 1X running buffer before adding 4 µL of sample containing 20 µg of the membrane protein or 2 µL of molecular weight marker. SDS-PAGE gels are run at a constant current of 10 mA.
4. During the run, PVDF membranes, cut just larger than the size of the gel, are immersed in 100% methanol for 15 s and then submerged and incubated in the blotting buffer with gentle agitation for at least 30 min. Complete wetting of the PVDF membrane is important to ensure proper blotting.

5. After the electrophoresis is finished, the gels are disassembled and equilibrated in blotting buffer for 10 min at room temperature.
6. Proteins on the gel are electrophoretically transferred to PVDF membranes. Extra thick blotting paper, pre-soaked in blotting buffer, is placed on to the platinum anode of a Trans-Blot SD Semi-Dry Electrophoretic Transfer Cell. Pre-wetted PVDF membrane is placed on the blot paper and then the equilibrated gel is placed on top of the PVDF membrane. Another pre-soaked blot paper is then placed on the gel. Air bubbles are carefully removed by rolling a glass pipette over the surface of the blot paper. The cathode is placed onto the stack (sandwich of blot paper-PVDF membrane-gel-blot paper) and the transfer is carried out at a constant current of 120 mA (2 mA/cm^2) for 30 min. Although the voltage limit is set to 15 V, to avoid excessive heating, transfer is usually completed below 10 V under these conditions.
7. After the transfer, the membrane is briefly rinsed with TBS-T and then incubated in 20 mL of blocking buffer for 1 h at room temperature on an orbital shaker to block non-specific binding sites of the PVDF membrane.
8. The blocking buffer is discarded and the membrane is incubated in a 1:5000 dilution of the anti-UGT1A antibody in blocking buffer for 1 h at room temperature on an orbital shaker (*see* **Note 5**).
9. After the membrane is washed three times for 5 min each with 50 mL of TBS-T, the membrane is incubated in 1:2000 dilution of the secondary antibody in blocking buffer for 1 h at room temperature on an orbital shaker.
10. The membrane is briefly rinsed with two changes of TBS-T, and then thoroughly washed four times for 10 min each with 50 mL of TBS-T.
11. During the final wash, ECL plus reagent (solution A and B), stored at 2 to 8°C, is equilibrated to room temperature. Solution A and B are mixed in a ratio of 40:1, or 4 mL of solution A and 0.1 mL of solution B (*see* **Note 6**).
12. The washed membrane is placed protein side up in the detection reagent and incubated for 5 min at room temperature, rotating by hand. The chemifluorescence signal is detected and quantified using the Typhoon 9400 variable mode imager (excitation; 457 nm, emission filter; 520BP40) and ImageQuant analysis software (Amersham Biosciences).
13. After the detection of chemifluorescence signal, the membrane is subsequently stripped and then reprobed with a polyclonal anticalnexin antibody. The membrane is incubated in 50 mL of stripping buffer for 30 min at 50°C with occasional agitation, extensively washed with distilled water until the lanes on the membrane become visible, and blocked again for 1 h in 20 mL of blocking buffer.
14. The membrane is then reprobed with anti-calnexin (1:100,000 in blocking buffer) by the same protocol as that for anti-UGT1A (*see* **Note 7**). A representative result of Western blotting is shown in **Fig. 1**.

3.4. Assay for SN-38 Glucuronidation

1. The assay mixture consists 40 µL of 500 m*M* Tris-HCl buffer (pH 7.4), 40 µL of 100 m*M* $MgCl_2$, 266 µL of distilled water, 4 µL of SN-38 solution, 10 µL of membrane fraction of COS-1 cells (100 µg protein), and 40 µL of 50 m*M* UDP–

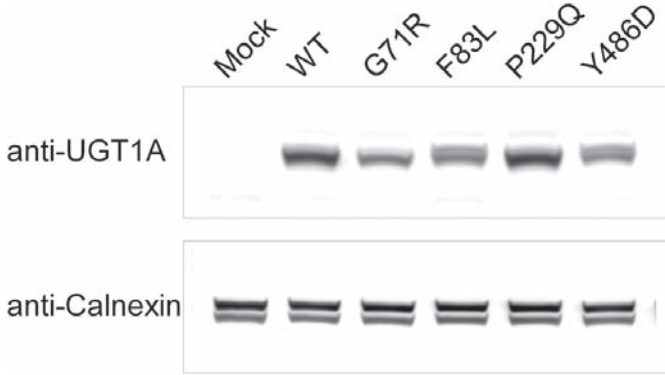


Fig. 1. Expression of wild-type (WT) and variant (G71R, F83L, P229Q, and Y486D) UGT1A1s in COS-1 cells. Aliquots (20 µg) of the membrane fractions were subjected to SDS-PAGE, electrophoretically transferred to a PVDF membrane, and immunochemically detected with a rabbit anti-human UGT1A antibody (1:5000) and ECL plus reagents. The membrane was subsequently stripped and reprobed with a rabbit anti-calnexin antibody (1:100,000) to show that the samples were evenly loaded. The decreased expression of G71R, F83L and Y486D UGT1A1 proteins was reproducibly shown in several transfection assays without a significant reduction in their mRNA levels (data not shown), suggesting that the G71R, F83L and Y486D UGT1A1 proteins are less stable or more rapidly degraded than the wild-type protein.

glucuronic acid. Prepare the mixture in an ice-cold 1.5-mL microtube by adding each component other than 50 m*M* UDP–glucuronic acid (*see* **Note 8** ***[11]***).

2. After preincubation in a shaking water bath at 37ºC for 1 min, the reaction is started by the addition of 40 µL of 50 m*M* UDP–glucuronic acid.
3. The mixture is incubated at 37ºC for 80 min, and the reaction is terminated with 100 µL of 10% (w/v) $HClO_4$ and vortexing.
4. After centrifugation at 12,000*g* for 10 min at 4ºC, the supernatant is filtered using a 0.45-µm PTFE membrane filter (Millipore, Bedford, MA) and subjected to HPLC analysis.
5. HPLC analysis is performed using a Shimadzu LC-10AD$_{VP}$ system (Kyoto, Japan) consisting of an SCL-10A$_{VP}$ controller, three LC-10AD$_{VP}$ pumps, a DGU-14A degasser, an SIL-10A$_{VP}$ auto injector with sample cooler, a CTO-10A$_{VP}$ column oven, an RF-10A$_{XL}$ fluorescence detector, and a C-R7A plus chromatopac integrator. The samples are cooled at 4°C, and 20-µL aliquots are injected into an Inertsil ODS-80A column (5 µm, 150 × 4.6 mm i.d., GL Sciences, Tokyo, Japan), which is kept at 40°C. The analyte is eluted isocratically with 50 m*M* KH_2PO_4 containing 3 m*M* sodium 1-octanesulfonate, pH 2.5/acetonitrile/methanol (72:22:6, v/v/v) at a flow rate of 1.0 mL/min. The excitation and emission wavelengths of the fluorescence detector are fixed at 370 and 425 nm, respectively (*see* **Note 9**).

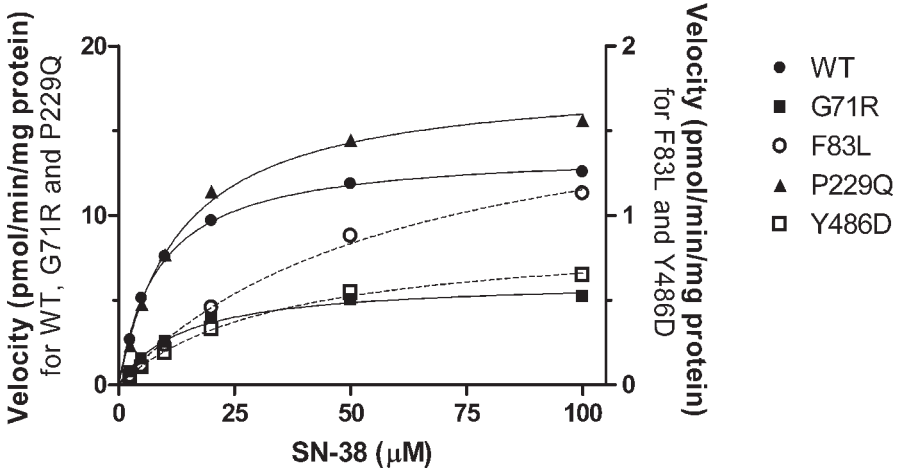


Fig. 2. Representative Michaelis–Menten kinetics of SN-38 glucuronidation by wild-type (WT) and variant (G71R, F83L, P229Q, and Y486D) UGT1A1s heterologously expressed in COS-1 cells. SN-38 glucuronidation by expressed UGT1A1s was assayed in the presence of the membrane fractions (100 μg) at a substrate concentration range between 2.5 and 100 μM. The solid and dashed lines indicate fitting of data to the Michaelis–Menten equation by nonlinear regression.

6. Kinetic parameters are calculated with Prism 4.00 (Graph Pad Software, Inc., San Diego, CA), using nonlinear regression of the Michaelis-Menten equation. Representative Michaelis-Menten kinetics of SN-38 glucuronidation by UGT1A1s are shown in **Fig. 2.**

4. Notes

1. To optimize the fidelity of PCR, concentrations of $MgCl_2$ and dNTP are lower than those used in the standard reaction conditions of 2.5 m*M* each.
2. The strategy initially is intended to subclone the UGT1A1 cDNA into pcDNA3.1(+). All the correct clones of pCR-UGT1A1/WT, however, contain the insert in a reverse orientation. Therefore pcDNA3.1(-) is applied here, which has a multiple cloning site in the opposite orientation to pcDNA3.1(+).
3. Mutation of the UGT1A1 cDNA is performed in the pCR 2.1 plasmid. The insert DNA is subsequently subcloned into the expression plasmid pcDNA3.1(–), thereby excluding possible mutations or PCR errors that may be introduced into the vector sequence.
4. Although Lipofectamine 2000 is can be used in the presence of FBS, transfection is conducted under serum-free conditions to obtain the maximal transfection efficiency.
5. Anti-UGT1A1 antibody is also commercially available from BD Gentest. However, we recommend anti-UGT1A antibody because of its higher affinity.

6. We preferably use ECL plus reagent for the detection of signals from Western blotting because this reagent is applicable for chemifluorescence detection as well as chemiluminescence detection. Fluorescence imaging by Typhoon 9400 offers a higher resolution and a wider linear dynamic range than chemiluminescence detection using a cooled-CCD camera.
7. It often is required that one ensure the SDS-PAGE samples are evenly loaded. Calnexin, an endoplasmic reticulum protein, is used for this purpose. Indeed, in some cases the expression level of a protein of interest (e.g., UGT1A1) is normalized by the amount of calnexin in each sample. In our experience, however, a Western blot of calnexin usually produce no detectable variation among the membrane fractions of COS-1 cells.
8. In the case of human microsome samples, pretreatment of the protein with alamethicin, a pore-forming peptide, is known to increase enzyme activity (***11***). This reagent is excluded from the SN-38 glucuronidation assay presented here because alamethicin has almost no effect on the glucuronidation activity of the COS-1 membrane fractions.
9. In the kinetic study, the amount of SN-38 glucuronide formed varies over a 300-fold range depending on the substrate concentrations and the membrane fractions of UGT1A1 variants. Therefore, it is important to adjust the dynamic range of the fluorescence detector appropriately. In this study, the dynamic range is achieved by setting the sensitivity to "medium" and the gain to "×1" of the Shimadzu RF-$10A_{XL}$ detector.

Acknowledgments

This study was supported in part by the Program for the Promotion of Fundamental Studies in Health Sciences (MPJ-6) of the Pharmaceuticals and Medical Devices Agency (PMDA). The authors thank Yakult Honsha Co. for generously donating SN-38 and SN-38 glucuronide, and Ms. Chie Knudsen for her secretarial assistance.

References

1. Ciotti, M., Basu, N., Brangi, M., and Owens, I. S. (1999) Glucuronidation of 7-ethyl-10-hydroxycamptothecin (SN-38) by the human UDP-glucuronosyltransferases encoded at the UGT1 locus. *Biochem. Biophys. Res. Commun.* **260,** 199–202.
2. Hanioka, N., Ozawa, S., Jinno, H., Ando, M., Saito, Y., and Sawada, J. (2001) Human liver UDP-glucuronosyltransferase isoforms involved in the glucuronidation of 7-ethyl-10-hydroxycamptothecin. *Xenobiotica.* **31,** 687–699.
3. Sai, K. Saeki, M., Saito, Y., Ozawa, S., Katori, N., Jinno, H., et al. (2004) UGT1A1 Haplotypes associated with reduced glucuronidation and increased serum bilirubin in irinotecan-administered Japanese cancer patients. *Clin. Pharmacol. Ther.* **75,** 501–515.

4. Ando, Y., Saka, H., Ando, M., et al. (2000) Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: a pharmacogenetic analysis. *Cancer Res.* **60,** 6921–6926.
5. Innocenti, F., Undevia, S. D., Iyer, L., et al. (2004) Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of irinotecan. *J. Clin. Oncol.* **22,** 1382–1388.
6. Sutomo, R., Laosombat, V., Sadewa, A. H., et al. (2002) Novel missense mutation of the UGT1A1 gene in Thai siblings with Gilbert's syndrome. *Pediatr. Int.* **44,** 427–432.
7. Saeki, M., Saito, Y., Jinno, H., et al. (2003) Comprehensive UGT1A1 genotyping in a Japanese population by pyrosequencing. *Clin. Chem.* **49,** 1182–1185.
8. Jinno, H., Tanaka-Kagawa, T., Hanioka, N., et al. (2003) Glucuronidation of 7-ethyl-10-hydroxycamptothecin (SN-38), an active metabolite of irinotecan (CPT-11), by human UGT1A1 variants, G71R, P229Q, and Y486D. *Drug Metab. Dispos.* **31,** 108–113.
9. Jinno, H., Saeki, M., Saito, Y., et al. (2003) Functional characterization of human UDP-glucuronosyltransferase 1A9 variant, D256N, found in Japanese cancer patients. *J. Pharmacol. Exp. Ther.* **306,** 688–693.
10. Jinno, H., Saeki, M., Tanaka-Kagawa, T., et al. (2003) Functional characterization of wild-type and variant (T202I and M59I) human UDP-glucuronosyltransferase 1A10. *Drug Metab. Dispos.* **31,** 528–532.
11. Soars, M. G., Ring. B. J., and Wrighton, S. A. (2003) The effect of incubation conditions on the enzyme kinetics of UDP-glucuronosyltransferases. *Drug Metab. Dispos.* **31,** 762–767.

3

Snapshot of the Allele-Specific Variation in Human Gene Expression

Hai Yan

Summary

The analysis of gene differential expression is complicated by the potentially subtle differences associated with alterations in a single allele as well as by variations between individuals that arise from environmental or physiological factors. To circumvent these analytic problems, a method, named allele-specific differential expression analysis, was developed to compare the relative expression levels of two alleles of the same gene within the same cellular sample. The studies of allele-specific expression revealed that differential expression is relatively common in the human population.

Key Words: Allele; expression; SNP; variation.

1. Introduction

Nucleotide polymorphisms can predetermine the complex genetic traits of individuals *(1)*. In particular, variations in the regulation of gene transcription are important in modulating the expression of the gene product in different tissues and at different stages of development *(2–4)*. Unfortunately, the regulatory sequences of most genes are not well characterized because the regulatory elements can be located thousands of bases from the transcription unit, and no reliable experimental approach exists to screen for regulatory functional polymorphisms. Moreover, the analysis of variation in gene expression is complicated by the potential small changes in expression associated with alterations in a single allele, as well as by potential variations between individuals that arise from trans-acting factors, the quality of messenger ribonucleic acid (mRNA) from different samples, and environmental influences *(3)*. To circumvent these analytical problems, the method of allele-specific differential expression has been used to measure the relative expression level of two alleles of one gene in the same cellular environment *(5,6)*. It can be postulated that, within

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

the same cellular environment, in the absence of *cis*-acting elements that can differentially affect the expression level of each copy of a gene, both alleles of the gene are equally expressed; in contrast, when an individual is heterozygous for any functional *cis*-acting polymorphisms that affect expression, the mRNA from each copy is expressed differentially. A single-nucleotide polymorphism (SNP) in the transcript can be used as a marker to distinguish between the transcripts derived from each allele. A fluorescent dideoxy terminator-based method was developed to distinguish and quantitate the mRNA products of alleles from normal individuals who were heterozygous for the marker SNP in the transcript of interest *(7)*.

2. Materials

2.1. Cell Culture

Lymphoblastoid cell lines representing genetically unrelated individuals from each of Centre du Etude Polymorphise Humain (CEPH) reference families were obtained from the National Institute of General Medical Sciences repository maintained by the Coriell Institute for Medical Research. Cells were grown in RPMI 1640 (Gibco/BRL, Bethesda, MD) supplemented with 10% fetal bovine serum (FBS, HyClone, Ogden, UT).

2.2. Genomic DNA, RNA Preparation, and Reverse Transcription

1. Genomic deoxyribonucleic acid (DNA) was extracted from cell lines with the DNeasy tissue kit (Qiagen, Valencia, CA).
2. mRNA was extracted with the QuickPrep™ Micro mRNA Purification Kit (Amersham Pharmacia, cat. no. 27-9255-01).
3. The SuperScript™ III First-Strand Synthesis System (Invitrogene 18080-051) was used for reverse transcription polymerase chain reaction (RT-PCR).

2.3. PCR

1. Primers were synthesized by MWG (High Point, NC).
2. 10 m*M* solution of dNTP (USB Corporation, cat. no. 77212).
3. Platinum Taq DNA Polymerase (Invitrogen, cat. no. 10966-034).
4. The AMPure Starter Kit (Agencourt cat. no. 000146, including 60 mL bottle of PCR clean up and a SPRI 96R magnet) was used for PCR product purification.

2.4. Snapshot Reaction

1. ABI PRISM® SNaPshot™ Multiplex Kit—1000 reactions (Applied Biosystems, cat. no. 4323161).
2. Sodium acetate, pH 4.6 (Applied Biosystems, cat. no. 400320).
3. Non-denatured 100% Ethanol.
4. Aluminum foil adhesive tape (Costar, cat. no. 6570).
5. 3M plastic PCR sealer (Bio-Rad, cat. no. 9101707).
6. HiDi formamide (Applied Biosystems, cat. no. 4311320).

3. Methods

To determine the allele-specific differential expression in normal individuals, lymphoblastoid cell lines representing genetically unrelated individuals were selected from CEPH references families. A quantitative assay for comparing the levels of relative expression of each allele in a sample. An SNP, which is in the transcript of interest and typically not the regulatory variant in the transcript, was used as a marker to discern the individual's paternal and maternal allele. The region surrounding the SNP in the transcript of interest was amplified by PCR. Relative expression of each allele was determined by using single-base extension, a fluorescent dideoxy terminator-based method, to distinguish the mRNA products of alleles from individuals who were heterozygous for the SNP. The reaction from single base extension was analyzed on an automatic sequencer. The peak heights of each of the two signals were measured, and the ratio of the signals was then converted to a ratio of fractional allelic expression.

3.1. Genomic DNA, mRNA Preparation, and RT

1. Genomic DNA was extracted from cell lines with the DNeasy tissue kit (Qiagen, Valencia, CA) according to manufacturer's instructions.
2. mRNA was prepared using QuickPrep™ Micro mRNA Purification Kit (Amersham Pharmacia, cat. no. 27-9255-01). In brief:
3. Dissolve 6.25×10^6 cells in cell pellet in 250 µL of extraction buffer
4. Add 500 µL of elution buffer to the extract, vortex, and Microfuge for 5 min at room temperature.
5. Transfer 1 mL of oligo-dT suspension to a 1.5-mL tube, spin 15 s to pellet oligo-dT matrix, and aspirate off the supernatant.
6. Transfer RNA supernatant to the oligo-dT matrix tube, vortex to mix, followed by end-to-end rotation for 10 min at room temperature.
7. Spin 15 s to pellet RNA-matrix and discard the supernatant.
8. Resuspend RNA-matrix in 1 mL of high salt buffer and spin 15 s to wash the RNA matrix and repeat this wash five times, wash once with 1 mL of low salt buffer.
9. Resuspend RNA-matrix in 0.5 mL of low salt buffer, transfer to a microspin column placed in a 1.5-mL screw cap tube and spin 5 s to get rid of buffer.
10. Prewarm elution buffer to 60°C; add 200 µL of prewarmed elution buffer to the column, spin 5 s to collect purified polyA+mRNA, and repeat this step with another 200 µL of elution buffer to the column.
11. Precipitate RNA with EtOH: 10 µL of glycogen, 40 µL of 2.5 *M* KCl, 1.0 mL of 95% EtOH, stand on dry ice for 10 min and warm to room temperature and spin for 10 min at room temperature.
12. Wash RNA pellet once with 70% EtOH and dry it at room temperature.
13. Resuspend RNA in total 54 µL of H_2O and keep it at 80°C until ready for RT reaction (*see* **Note 1**).

14. RT: 200 ng of poly(A)+ mRNA samples were reverse transcribed by using the SuperScript™ III First-Strand Synthesis System (Invitrogen, cat. no. 18080-051; available at: http://invitrogen.com/Content/sfs/manuals/superscriptIIIfirststrand_pps.pdf). No RT control samples were made in parallel during first-strand complimentary DNA (cDNA)–synthesis steps. Resulting samples were diluted 10-fold for PCR amplification.

3.2. PCR Amplification of the Regions Containing the Marker SNP

1. Using Celera SNP database, we selected a SNP in the transcript of interest as the marker to distinguish the mRNA products of alleles from individuals who were heterozygous for the SNP (*see* **Note 2**).
2. Two sets of primers used for amplification of a 200- to 300-bp region containing a marker SNP were designed by Primer 3.0 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi; melting temperature of 59–63°C, 30–70% CG content; 18–26 bps primer length). One pair of primers was designed for the genomic DNA amplification. Another pair of primers was used for cDNA amplification (*see* **Note 3**).

3.2.1. PCR Conditions

Make a master mix by setting up the following for each 20-µL reaction:

H_2O	12.8 µL
10x PCR buffer	2 µL
dNTP (10 m*M*)	1 µL
DMSO	1.2 µL
Forward primer (50 µ*M*)	0.4 µL
Reverse primer (50 µ*M*)	0.4 µL
Add Platinum Taq, 5 u/µL	0.2 µL
DNA (genomic DNA 5 ng/µL or cDNA)	2 µL

For each sample, we set up seven replicates for both genomic and cDNA amplifications. The reactions should be run as follows (*see* **Note 4**): 1 cycle of 94°C, 2 min; 35 cycles of 94°C, 15 s; 55°C, 15 s; 70°C, 15 s; and 1 cycle of 70°C, 5 min.

3.2.2. PCR Products Purification

1. Thoroughly mix AMPure Binding Solution by swirling and inverting until it appears homogenous and consistent in color.
2. Add 18 µL of binding solution to 10 µL of PCR and mix by pipetting up and down five times and stand the mix for 5 min.
3. Place plate onto SPRIPlate 96 Magnet, and allow beads to bind to the magnet for 5 min or until the solution appears clear.
4. Aspirate 20 µL of solution waste from each well of the plate and discard.
5. Spin plate inverted on a paper towel at 30*g* for 30 s to remove any remaining waste.

6. Dispense 200 µL of 70% ethanol into each well of the plate (tips point toward bottom of wells to wash all waste from inside the wells) and allow plate to stand for at least 30 s.
7. Spin plate inverted on a paper towel at 30*g* for 30 s to remove ethanol (plate remains on magnet during this step) and repeat ethanol wash and removal.
8. Place plate on bench top to dry for 15 min (do not remove plate from magnet as samples become powdery when dry and can easily be blown out of their wells.
9. Add 30 µL of TE to each well then remove plate from magnet and spin down at a speed of 82.3*g*; wait 30 min before proceeding to next step to allow DNA to elute from beads.
10. Place plate on magnet and allow solution to separate for 10 min or until solution appears clear; while waiting for solution to separate, it is possible to draw off 3 µL of purified product to run on an agarose gel without effecting final product.
11. Transfer 20 µL of purified product to a new 96-well plate and use the purified product for subsequent sequencing steps.

3.3. Snapshot Single-Base Extension

3.3.1. Single-Base Extension Primers

Extension primers: each extension primer ends at the nucleotide adjacent to the SNP of interest. Different sizes of forward and reverse extension primers can be applied together in the same reaction to provide a pair of confirmative data. For example, two pair of primers can be designed for the following SNP and its adjacent sequence:

tga gccagggacg tgctgggaaa gcccaagcc**C/T** cgggagaaga tgccggccat cctggtcgcc agt
Pair one: 26 bp forward (5'-3' agggacg tgctgggaaa gcccaagcc) and
31 bp reverse primers (5'-3't ggcgaccagg atggccggca tcttctcccg)
Pair two: 26 bp reverse (5'-3' accagg atggccggca tcttctcccg) and
31 bp forward primers (5'-3' ga gccagggacg tgctgggaaa gcccaagcc)

Each pair of primers should be tested on a couple of heterozygous samples to determine which pair gives better signal to noise ratio (*see* **Note 5**).

3.3.2. Prepare Snapshot Reaction

1. Thaw the reagents in ABI PRISM SNaPshot Multiplex Kit on ice just before the reaction and keep the reaction mix in dark and on ice.
2. Set up reactions in single 0.2-mL PCR tubes or 96-well PCR plates.

Snapshot Multiplex Ready reaction Mix	2.5 µL
Purified PCR product	1.5 µL
Extension primers (1 µ*M*, *see* **Note 6**)	0.5 µL forward plus 0.5 µL reverse
DI water	0.5 µL
Total	5 µL

3.3.3. Thermal Cycling the Extension Reaction

Place the plate on a thermal cycler and repeat the following for 25 cycles: rapid thermal ramp to 96°C; 96°C for 10 s; rapid thermal ramp to 50°C; 50°C for 5 s; rapid thermal ramp to 60°C; 60°C for 30 s and hold at 4°C.

3.3.4. Purification of the Snapshot Reaction

1. Prepare the following precipitation mixture:

3 *M* NaOAc, pH 4.6 (PE p/n 400320)	0.75 µL
Nondenatured 100% ethanol	17.5 µL
dH_2O	1.75 µL

2. Add 20-µL precipitation mixtures to 5-µL Snapshot reactions and mix three times.
3. Cover with aluminum foil adhesive tape (Costar, cat. no. 6570) and seal wells with the edge of a 3M plastic "PCR plate sealer" device.
4. Spin for 1 s at 6200*g*; store at room temperature for 15 min (no longer than 4 h); spin 30 min at 6200*g*; invert onto paper towel, spin inverted for 60 s at 20*g*.
5. Add 25 µL of 70% ethanol/30% water and spin for 5 min at 6200*g*.
6. Invert onto paper towel, spin inverted for 60 s at 20*g*; spin in speed vacuum for 8 min, 35°C (remember to set "heat time" on Speed Vacuum to 5 min).
7. Add 30 µL of HiDi formamide, cover with aluminum foil adhesive tape; spin for 1 s at 6200*g*; triturate five times or allow pellets to resuspend for at least 4 h at 4°C; transfer contents to 96-well plates and store at –20°C.

3.3.5. Size Separation of the Extended Primers by Electrophoresis

Size separate the extended primers by electrophoresis through the Spectru Medix SCE9610 Genetic Analysis system (*see* **Note 7**) and run at 0.5 to 2 KV (*see* **Note 8**) for 30 min. We determined the peaks of dye intensities corresponding to extension of single base extension primers by inspecting output from the SpectruMedix SCE9610 Genetic Analysis System after background subtraction and color separation.

3.3.6. Genomic DNA Snapshot

First, the extension primers for the SNP of interest were tested on seven replicates of individual genomic DNA samples to determine if the individual is heterozygous for the SNP. Before subsequent statistical analyses, obvious technical failures or statistical outliers were eliminated. For the individual who is heterozygous for the SNP, the ratio of the magnitude of the peak heights from the two nucleotides at the same sequencing position was calculated. The average ratio from the seven replicates was then derived.

3.3.7. cDNA Snapshot

For assays performed on the cDNA, the fractional allelic experiment for each sample was also determined through seven replicates and obvious technical failures or statistical outliers were eliminated. The allelic ratio of expression based on the magnitudes of SNP peak heights were determined. Each individual's average allelic differential expression ratio value can be normalized based on the average allelic ratio derived from the same individual's genomic template.

4. Notes

1. To avoid genomic DNA contamination of the RT-PCR, mRNA was treated with DNase.
2. SNPs, which are present with a relatively high frequency in population were chosen for this study.
3. For cDNA PCR amplification, we selected primers spanning large introns to avoid genomic DNA contamination.
4. For genes expressed at low levels, the potential nonhomogenous distribution of DNA templates might cause variations within aliquots of the sample and result in biased allelic ratio of the gene transcripts. Quantitative real-time PCR is necessary to reveal the samples that have very low copy numbers of DNA templates. Data derived from the samples were excluded for the subsequent analysis.
5. The extension primers should meet the following conditions: 1) the assay was robust and 2) each primer gave a single dye peak in genomic DNA.
6. The concentration of extension primers might affect the signal to noise ratio. If the noise is high, decrease the concentration; on the contrary, increase the concentration.
7. An ABI sequencer can be the alternative of SpectruMedix SCE9610 Genetic Analysis System.
8. The Snapshot samples were run on SpectruMedix SCE9610 Genetic Analysis System at 0.5 kV for 30 min. If the signal is low, the run can be adjusted to 2 kV.

Acknowledgments

The author would like to thank Dr. Victor Veculescu at Johns Hopkins for development of the sequencing protocols for using SpectruMedix SCE9610 Genetic Analysis System and Daniel Broderick for his thoughtful edits.

References

1. Hamilton, B. A. (2002) Variations in abundance: genome-wide responses to genetic variation and vice versa. *Genome Biol.* **3,** reviews 1029.
2. Wray, G. A., Hahn, M. W., Abouheif, E., et al. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20,** 1377–1419.

3. Yan, H. and Zhou, W. (2004) Allelic variations in gene expression. *Curr. Opin. Oncol*. **16,** 39–43.
4. Knight, J. C. (2004) Allele-specific gene expression uncovered. Trends Genet., **20,** 113–116.
5. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002) Allelic variation in human gene expression. *Science* **297,** 1143.
6. Cowles, C. R., Hirschhorn, J. N., Altshuler, D., and Lander, E. S. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet*. **32,** 432–437.
7. Matyas, G., Giunta, C., Steinmann, B., Hossle, J. P., and Hellwig, R. (2002) Quantification of single nucleotide polymorphisms: a novel method that combines primer extension assay and capillary electrophoresis. *Hum. Mutat*. **19,** 58–68.

4

Genome-Wide Analysis of Allele-Specific Gene Expression Using Oligo Microarrays

Maxwell P. Lee

Summary

Human variation is largely caused by deoxyribonucleic acid polymorphism and difference in gene expression. Common disease/common variant hypotheses suggest that quantitative differences among different alleles may be the basis for complex diseases. Quantitative difference in gene expression between alleles may affect most complex diseases. We have developed a gene chip-based method to quantitatively examine allele-specific gene expression of 1063 transcribed single-nucleotide polymorphisms using Affymetrix HuSNP oligo arrays. Among the 602 genes that were heterozygous and expressed in kidney or liver tissues from seven individuals, 326 (54%) showed preferential expression of one allele in at least one individual. The genes that showed allele-specific expression are distributed throughout the genome. We showed that variation of gene expression between alleles is common and that this variation may contribute to human variation. Our studies demonstrate the feasibility to perform genome-wide analysis of allele-specific gene expression.

Key Words: Allele-specific expression; SNP; gene chip; genotyping; genetic variation; complex disease.

1. Introduction

Polymorphism and variation in gene expression provide the genetic basis for human variation. Mendelian diseases are caused by mutations in a single gene or a few genes. To date, mutations in more than 2000 genes have been identified (http://www.ncbi.nih.gov/entrez/query.fcgi?db=OMIM). Most of these mutations change the protein structure and function. Increasing efforts have been made toward understanding the genetic basis of common complex diseases. It is commonly believed that the complex diseases are caused by com-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

bination of common single-nucleotide polymorphisms (SNPs), each of which contributes quantitatively to the diseases. Most of the efforts so far have been focused on nonsynonymous SNPs. However, most of the SNPs in the genome are not nonsynonymous SNPs. They instead are synonymous SNPs, SNPs in untranslated region, intronic SNPs, or intergenic SNPs. These SNPs can affect complex diseases through their effects on gene expression.

Currently, there are more than 2 million SNPs deposited in GenBank (http://www.ncbi.nih.gov/SNP/). It is daunting task to perform association studies for all those SNPs. Several initiatives have been taken to prioritize a subset of those SNPs for association study of various types of diseases. Those include Haplotype Map Project and candidate SNP approaches. A SNP outside coding region may affect gene product quantitatively by altering gene expression. This type of SNP should show difference in gene expression between the two alleles of an individual. Identifying this class of SNPs could have significant impact in our efforts to identify genes that are associated with complex diseases. Several recent studies have shown that allelic variation in gene expression is common in the human genome ***(1–3)***. Also, variation in allelic gene expression was shown to be transmitted by Mendelian inheritance ***(1)***. To address the feasibility to analyze allele-specific gene expression at genome-wide level, we modified an existing genotyping technology, the Affymetrix HuSNP chip system, to analyze allele-specific gene expression ***(2)***.

The HuSNP chip was designed for simultaneous typing of 1494 SNPs of the human genome ***(4)***. It has been applied successfully to study loss of heterozygosity in human cancer ***(5)***. The HuSNP chip contains 16 probes for each SNP locus (details can be found in **Subheading 2**), with four matching perfectly to allele A and the other four matching perfectly to allele B. The other eight probes differ from the first eight probes by having one mismatched base in the center of the probe. In this report, we summarize the method that we used to perform both genotyping and allele-specific gene expression using HuSNP chips. Our studies demonstrate that the HuSNP chip system is a reliable way to simultaneously measure allele-specific gene expression for hundreds of genes.

2. Materials

2.1. Fetal Tissues

Fetal tissues were obtained from the Birth Defects Research Laboratory, University of Washington. The tissues were snap-frozen after surgery and were stored in liquid nitrogen. Kidney and liver tissues from seven individuals, five male and two female, were used in this study. The ages of the fetuses ranged from 78 to 103 d.

2.2. DNA Isolation

Genomic DNA was isolated using the QIAamp DNA mini kit (Qiagen, Inc., Valencia, CA).

1. Tissues of 25 mg were cut into small pieces and were placed in a 1.5-mL tube.
2. The tissues were mixed with 180 µL of Buffer ATL and 20 µL of proteinase K and incubated at 56°C until tissue was completely lysed, using shaking water bath to ensure mixing of sample.
3. The tissues were mixed with 200 µL of Buffer AL by vortexing for 15 s and then incubated at 70°C for 10 min, followed by addition of 200 µL of ethanol.
4. Tissue mixture was applied to QIAmp spin column without wetting the rim and was spun at 6000*g* for 1 min.
5. The samples were washed with 500 µL of Buffer AW1 (with ETOH added) and then 500 µL of Buffer AW2.
6. Genomic DNAs were eluted with 200 µL of Buffer AE or distilled H_2O.

2.3. RNA Isolation

1. RNAs were isolated from fetal tissues using RNAzol[B] (Tel-Test, Inc., Friendswood, TX) according to the manufacturer's protocol.
2. Tissues of 50 mg were homogenized in 4 mL of RNAzol[B] and were mixed thoroughly with 0.8 mL of $CHCl_3$ and incubated on ice for 5 min.
3. The samples were spun at 9300*g* for 10 min at 4°C.
4. Aqueous phase (2 mL) were collected and mixed with 2 mL of isopropanol and incubated at room temperature for 5 min.
5. The samples were spun at 13,400*g* for 10 min at 4°C and washed with 4 mL of cold 75% EtOH.
6. RNAs were suspended in 50 µL of DEPC-treated H_2O containing 1 µL of 40 U/µL RNase inhibitor and were stored at –70°C.
7. Poly-A RNAs were isolated using the Micro-Fast Track kit (Invitrogen Corp., Carlsbad, CA). We mixed 20 µL of RNA with 1 mL of lysis buffer.
8. The mixture was incubated at 65°C for 5 min and chilled on ice for 1 min.
9. The samples were transferred back to room temperature and mixed with 63 µL of 5 *M* NaCl.
10. The samples were mixed with oligo (dT) cellulose for 20 min at room temperature and spun at 133*g* for 5 min at room temperature. The pellet was suspended in 1.3 mL of binding buffer. The above step was repeated until OD_{260} in the supernatant is less than 0.05.
11. The pellet was washed with 500 µL of Low Salt Wash buffer twice and eluted with 100 µL of elution buffer.
12. Poly A RNA was precipitated with NaAc and EtOH and resuspended in 20 µL of DEPC-treated H_2O, containing 1 µL of 40 U/µL RNase inhibitor and was stored at –70°C.

2.4. cDNA Synthesis

1. Complimentary deoxyribonucleic acid (cDNA) synthesis was conducted in 50-µL reaction. Poly-A RNA (60 ng) was mixed with 1 µg of oligo dT, 3 µg of random hexamer, and DEPC-treated water to a final volume of 34 µL.
2. The sample was incubated at 70°C for 10 min and put back on ice.
3. We added 10 µL of 5X buffer, 4 µL of dNTP (each at 2.5 m*M*), 1 µL of 20 U/µL RNase inhibitor, and 1.5 µL of 2.5 U/µL AMV reverse transcriptase (NEB, Beverly, MA) to RNA/primer mixture.
4. The mixture was incubated at 42°C for 1 h and stored cDNA at –70°C.

3. Methods

3.1. HuSNP Experiments

1. The HuSNP experiments were conducted according to the GeneChip HuSNP Mapping Assay Manual (P/N 700308, Affymetrix, INC., Santa Clara, CA). Genomic DNA (120 ng) or cDNA (6 ng) was used for each set of 24 multiplex polymerase chain reactions (PCRs; *see* **Notes 1** and **2**).
2. We prepared PCR I mixture by adding 35 µL of 10X buffer, 70 µL of 25 m*M* $MgCl_2$, 70 µL of 2.5 m*M* dNTP, and 7 µL of 5 U/µL AmpliTaq gold DNA polymerase (Applied Biosystems, Foster City, CA).
3. We took 223 µL of PCR I mixture and added 34 µL of 4 ng/µL DNA (or 34 µL of 0.2 ng/µL cDNA).
4. We dispensed 9.5 µL of the mixture into each well of a 96-well plate.
5. We added 3 µL of multiplex primers (24 primer sets for each DNA) to 96-well plate, sealed the plate with a film, did a quick spin, and put the plate on a thermal cycler.
6. The PCR condition had one cycle of 95°C for 5 min, 30 cycles of 95°C for 30 s, 52°C approx 57.8°C for 55 s (starting from 52°C and increasing the temperature by 0.2°C per cycle), 72°C for 30 s, and followed by five cycles 95°C for 30 s, 58°C for 55 s, 72°C for 30 s, and extension at 72°C at 7 min.
7. PCR II mixture contained 187.5 µL of H_2O, 62.5 µL of 10X buffer, 100 µL of 25 m*M* $MgCl_2$, 100 µL of 2.5 m*M* dNTP, 50 µL of 10 µ*M* biotin-T7 primer, 50 µL of 10 µ*M* biotin-T3 primer, and 12.5 µL of U/µL AmpliTaq gold DNA polymerase.
8. We aliquoted 22.5 µL of PCR II mixture into the 96-well plate, added 2.5 µL of 1:1000 dilution of PCR I product (from serial dilution of 3X 1:10), sealed the plate with a film, did a quick spin, and put on the thermal cycler.
9. PCR condition II contained one cycle of 95°C for 8 min, 40 cycles of 95°C for 30 s, 55°C for 90 s, 72°C for 30 s, and extension at 72°C for 7 min.
10. We used 1 µL of the PCR II product to check PCR reaction by gel electrophoresis (*see* **Note 3**).
11. We took 23 µL of PCR II products and combined those (12 wells per tube), put the PCR products into a Microcon-10 concentrator, and spun at 13,000*g* for 10 min.
12. We then inverted the concentrator into a fresh tube to collect samples and spun at 3000*g* for 3 min. You should get about 30 µL from each tube.
13. The samples were stored at –20°C.

14. Hybridization solution: mix 4 µL of H_2O, 81 µL of 5 *M* TMAC, 1.4 µL of control oligo B1, 1.4 µL of 1 *M* Tris-HCl, pH 7.8, 1.4 µL of 1% Tween-20, 1.4 µL of 0.5 *M* EDTA, 1.4 µL of 10 mg/mL sonicated Salmon Sperm DNA, 13.5 µL of 50X Denhardt's solution, and 30 µL of concentrated PCR II products.
15. The hybridization mixture was denatured at 95°C for 5 min and was added to a HuSNP chip.
16. The hybridization was carried out at 44°C for 16 h.
17. Washing buffer A: 140 mL of H_2O, 60 mL of 20X SSPE, and 0.2 mL of 10% Triton X-100.
18. Washing buffer B: 160 mL of H_2O, 40 mL of 20X SSPE, and 0.2 mL of 10% Triton X-100.
19. Staining solution: 305 µL of H_2O, 150 µL of 20X SSPE, 10 µL of 50X Denhardt's solution, 5 µL of 1% Tween-20, 12.5 µL of 1 mg/mL SAPE (Molecular Probes, Eugene, OR), and 5 µL of 0.5 mg/mL Biotinylated anti-streptavidin (Vector Laboratories, Burlingame, CA).
20. The HuSNP chip was washed and stained with the following Fluidics Protocol.

Washing A1 temperature (°C)	25
Number of wash A1 cycles	2
Mixes per wash A1 cycle	2
Washing B temperature (°C)	35
Number of wash B cycles	6
Mixes per wash B cycle	5
Staining time (s)	1800
Staining temperature (°C)	25
Washing A2 temperature (°C)	25
Number of wash A2 cycles	6
Mixes per wash A2 cycle	4
Holding temperature (°C)	25

21. The chip was then scanned in a HP GeneArray Scanner (Affymetrix, Inc., Santa Clara, CA). Genotyping calls were made using the Affymetrix MicroArray Suite (MAS) software version 4.0 (*see* **Note 4**). Allele-specific gene expression was analyzed by the method described in **Subheading 3.2.**

3.2. Computational Analysis of HuSNP Data

We mapped SNPs in the transcribed region using the annotation in dbSNP (http://www.ncbi.nih.gov/SNP/) and Blast search. The criteria for Blast search were: 1) at least two EST hits; 2) E-value $< 10^{-10}$; and 3) alignment > 40 bp. We were able to map 1063 SNPs to the transcribed regions of genes.

We extracted the intensity values for each probe from the .CEL files generated by Affymetrix MAS 4.0. The .CEL files contain the fluorescent intensity values for each of the probes. HuSNP chip contains minimally 16 probes for each SNP locus. Four of the 16 probes match perfectly to allele A, four to allele B, four have one mismatch to allele A, and the other four have one mismatch to allele B. Allele A and allele B represent the two alleles of the SNP. Each probe

contains 20 nucleotides. Affymetrix defines a mini-block as a group of four probes that include a perfect match probe for allele A (PMA), a mismatch probe for allele A (MMA), a perfect match probe for allele B (PMB), and a mismatch probe for allele B (MMB). The mismatch probe has one mismatch base in the center of the probe. There are four miniblocks for each SNP, with the center nucleotide of the probe corresponds the base at –4, –1, 0, and 1 of the original SNP sequence. Ninety-five SNPs have an additional probe with the center base at +4 position of the SNP. The value for each probe pair was computed by subtracting the mismatch intensity from the perfect match intensity. A *t*-test was used to calculate a *p* value for the presence of signal (intensity greater than 0) for each allele of each SNP. We considered a signal to be present if at least one allele had signal ($p < 0.01$, *t*-test). For those SNP with signal, we set (PMA – MMA) = 50 if (PMA – MMA) is less than 50 for each miniblock. Similarly, baseline for allele B was set at 50. Fraction of the A allele, defined as f = (PMA-MMA)/(PMA-MMA + PMB-MMB), was computed for each mini-block, and the mean of the fractions among the four mini-blocks was computed for each SNP. The ratio of allele A/allele B can be computed from f/(1 – f). Two scans, scan A and scan B, are taken for each chip. Generally, we used the intensity values from scan A. We used the intensity values from scan B if scan A doesn't have a signal defined by *t*-test. The ratio of two alleles in cDNA was further normalized by the ratio of genomic DNAs for the SNP. Among the 602 SNPs analyzed in our studies, 39 had at least five heterozygous fetuses. We computed the 95% confidence interval for the allelic ratio of genomic DNA for each of these 39 SNPs, and the average confidence interval was between 0.5 and 2.0. This value was used to select those genes that show significant difference in the expression between the two alleles (*see* **Notes 5** and **6**).

4. Notes

1. It is essential to start with high molecular weight genomic DNA. The quality of the genomic DNA was analyzed by 1% agarose gel electrophoresis with and without a restriction enzyme digestion. There should be only high molecular weight DNA before restriction enzyme digestion and generation of a mixture of variable sizes of DNA fragments after digestion.
2. The protocol used 120 ng genomic DNA. It is generally believed that 2 to 5% of genomic DNA contains transcribed sequences. Therefore we used 6 ng of cDNA made from poly-A RNA. It is also possible to do the allele-specific gene expression using cDNA made from total RNA. Contamination of genomic DNA will increase the call for bi-allelic expression. However, genes that displayed preferential expression of one allele are not owing to any trace amount of genomic DNA contamination.
3. We always check the PCR II by a 4% agarose gel before hybridization. A typical gel picture is shown in **Fig. 1**. The expected PCR products are around 90 to 100 bp in all 24 multiplex PCRs.

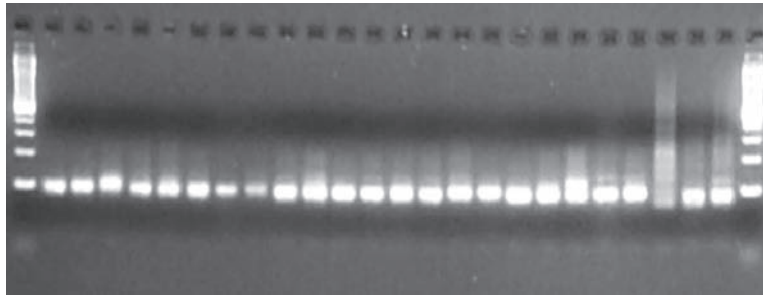


Fig. 1. Evaluation of PCR product by a gel electrophoresis. Quality of the PCR products was analyzed by a 4% agarose gel. The expected PCR products are approx 90–100 bp. The first and last lanes contain 100-bp ladder as size markers. The 24 lanes between the size markers contain 24 multiplex PCR reactions. The PCR reactions 1–21 contain 1494 SNPs and the rest of the PCR reactions have the controls.

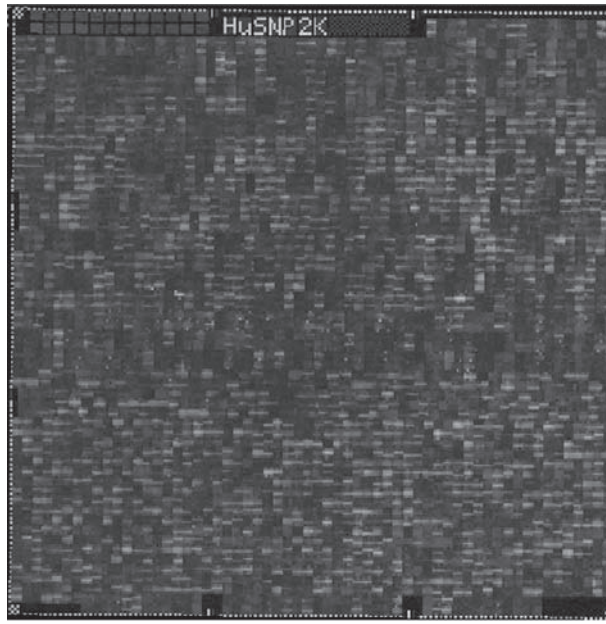


Fig. 2. Scan image of a HuSNP chip. The scan image is from one of the genomic DNA HuSNP experiment. HuSNP 2k label should be clear.

4. We used Affymetrix MAS 4.0 in this study. A typical scan image is shown in **Fig. 2**. You should see a sharp image with HuSNP 2K label and signals evenly distributed across the chip. A report for the image is presented in **Table 1**. A1 to A21 correspond to 1 to 21 of the 24 multiplex PCRs (**Fig. 1**, lanes 1–21), which

Table 1
A Report From HuSNP Array

Pool	%Pass	%A	%AB	%B	%AB_A	%AB_B	%No signal
A01	79.8	17	36.2	24.5	2.1	0	20.2
A02	72.9	25.9	28.2	14.1	4.7	0	27.1
A03	86.7	32.5	20.5	30.1	2.4	1.2	13.3
A04	65.8	27.6	10.5	25	1.3	1.3	34.2
A05	91.7	27.8	13.9	47.2	2.8	0	8.3
A06	63.4	26.8	15.5	19.7	1.4	0	36.6
A07	60	16	21.3	18.7	2.7	1.3	40
A08	76.3	23.8	26.3	25	0	1.3	23.8
A09	79.3	23.2	31.7	24.4	0	0	20.7
A10	72.9	21.2	15.3	32.9	0	3.5	27.1
A11	77.9	25.6	18.6	31.4	2.3	0	22.1
A12	87	35.1	27.3	24.7	0	0	13
A13	85.2	26.1	25	33	1.1	0	14.8
A14	80	30	22.9	24.3	1.4	1.4	20
A15	78.1	23.3	26	26	0	2.7	21.9
A16	71.1	22.4	26.3	21.1	0	1.3	28.9
A17	69.2	16.7	28.2	20.5	3.8	0	30.8
A18	62	6	18	34	4	0	38
A19	78.4	31.4	17.6	27.5	0	2	21.6
A20	59	15.4	20.5	23.1	0	0	41
A21	69.2	20.5	17.9	30.8	0	0	30.8
Total	75	23.8	23	25.9	1.5	0.8	25

An example of a report from HuSNP experiment is presented here. The report in **Table 1** and the image in **Fig. 2** are from the same HuSNP array.

contain 1494 SNPs. The calling rate varies among different multiplex PCR reactions. We have used MAS 5.0 in recent studies (unpublished data). MAS 5.0 outputs RAS values that are similar to the fraction generated by the method described in this chapter.

5. The reproducibility of the HuSNP system can be assessed by correlation between duplicated experiments. To evaluate concordance between two duplicate experiments, we computed the Pearson correlation coefficient between the two experiments using the mean intensity of the probe pairs from each allele of a SNP. Pearson correlation coefficient is 0.98 to 0.99 for genomic DNA and 0.88 to 0.96 for cDNA *(2)*.
6. The method described here can also apply to the study using Affymetrix 10k chip and 100k chip. The new high-density oligo arrays use GCOS software, which outputs RAS values in addition to genotype call.

Acknowledgments

I would like to thank Dr. H. Shuen Lo and Sheryl Gere for technical assistance and Drs. Howard Yang and Ying Hu for bioinformatics supports.

References

1. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002) Allelic variation in human gene expression. *Science* **297,** 1143.
2. Lo H. S., Wang Z., Hu Y., Yang H. H., Gere S., Buetow K. H., and Lee M. P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res*. **13,** 1855–1862.
3. Pastinen T., Sladek R., Gurd S., et al. (2004) A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics*. **16,** 184–193.
4. Mei R., Galipeau P. C., Prass C., et al. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res*. **10,** 1126–1137.
5. Lindblad-Toh, K., Tanenbaum, D. M., Daly, M. J., et al. (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18,** 1001–1005.

5

HaploChIP

An In Vivo Assay

Julian Charles Knight

Summary

The characterization of protein–deoxyribonucleic acid (DNA) interactions occurring at an allele-specific level is important to resolving the functional consequences of genetic variation in non-coding DNA for gene expression and regulation. The approach of haplotype-specific chromatin immunoprecipitation (i.e., haploChIP) resolves in living cells relative protein–DNA binding to a particular allele through immunoprecipitation of proteins crosslinked to DNA. Single-nucleotide polymorphisms present in a heterozygous form are used as markers to differentiate allelic origin. This in turn allows resolution of specific haplotypes showing differences in relative protein occupancy. The haploChIP approach allows testing of in vitro hypotheses that a transcription factor protein shows haplotype specific occupancy. In addition, the haploChIP approach allows screening of haplotypes for differences in relative gene expression by immunoprecipitation using antibodies to phosphorylated Pol II.

Key Words: Transcription; haploChIP; allele-specific; chromatin immunoprecipitation; gene expression; polymorphism; RNA polymerase II.

1. Introduction

The approach of haplotype-specific chromatin immunoprecipitation (haploChIP) was developed to aid the functional characterization of genetic variation in noncoding deoxyribonucleic acid (DNA ***[1]***). Such variation, notably single-nucleotide polymorphisms (SNPs), is common in the genome and has been implicated in determining our susceptibility to many complex diseases, notably infectious, autoimmune, inflammatory, and malignant conditions ***(2)***. In coding DNA, the consequences of variation for protein structure and function are amenable to functional prediction and testing ***(3)***. In contrast, variation in noncoding DNA, which modulates gene regulation and expression, is difficult to resolve with most variation having no functional impor-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

tance *(4)*. One approach is to use transcribed polymorphisms to determine the allelic origin of transcripts and thus relative allele-specific expression *(5)*. For many genes and most haplotypes, such exonic polymorphisms are absent, which precludes such analysis. The haploChIP approach initially was developed to address this problem by immunoprecipitating in living cells phosphorylated Pol II bound to a gene to provide a surrogate measure of relative gene expression *(1)*. SNPs in a heterozygous state are used to determine the relative allelic abundance of immunoprecipitated DNA and, thus, Pol II loading between the two alleles (**Fig. 1A**). This considerably broadens the scope for analysis as any SNP within approx 2 kb of a gene can be used as a marker. The approach does however require knowledge of underlying haplotypic structure to enable full interpretation of relative allelic differences between different individuals.

A natural progression of the haploChIP approach is to test specific hypotheses relating to mechanisms of gene regulation for individual DNA-binding proteins. Thus, if in vitro analysis predicts allelic differences in DNA binding, this can be tested in vivo using the haploChIP approach by use of specific antibodies for that transcription factor *(6)*. HaploChIP uses the technique of chromatin immunoprecipitation *(7)* to crosslink protein to DNA in cells using formaldehyde from which the nuclear material is extracted, sonicated and subject to cesium chloride centrifugation. After immunoprecipitation using specific antibodies, protein–DNA crosslinks are reversed and protein digested away, leading to a pool of DNA fragments that can be analyzed by gene-specific polymerase chain reaction (PCR) and primer extension to allow allelic discrimination (**Fig. 1B**). Accurate and specific allele-specific quantification of the products of primer extension for a specific SNP can be achieved by a number of approaches, notably matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF-MS *[8,9]*).

2. Materials

2.1. Preparation of Crosslinked Chromatin

2.1.1. Cell Culture

1. RPMI 1640 (Sigma-Aldrich, Gillingham, Dorset, UK) supplemented with 2 m*M* glutamine (Sigma) and 10% fetal bovine serum (FBS; Sigma).
2. Phorbol 12-myristate 13-acetate (Sigma) dissolved at 1 m*M* in dimethyl sulfoxide and stored in single-use aliquots at –20°C.
3. Ionomycin (Sigma) dissolved at 1 m*M* in dimethyl sulfoxide and stored in single-use aliquots at –20°C.

2.1.2. Formaldehyde Crosslinking

1. Formaldehyde crosslinking buffer (10X): 100 m*M* NaCl (Sigma), 1 m*M* ethylenediamine tetraacetic acid (EDTA), pH 8.0 (Sigma), 0.5 m*M* m*M* *N*-hydroxy-

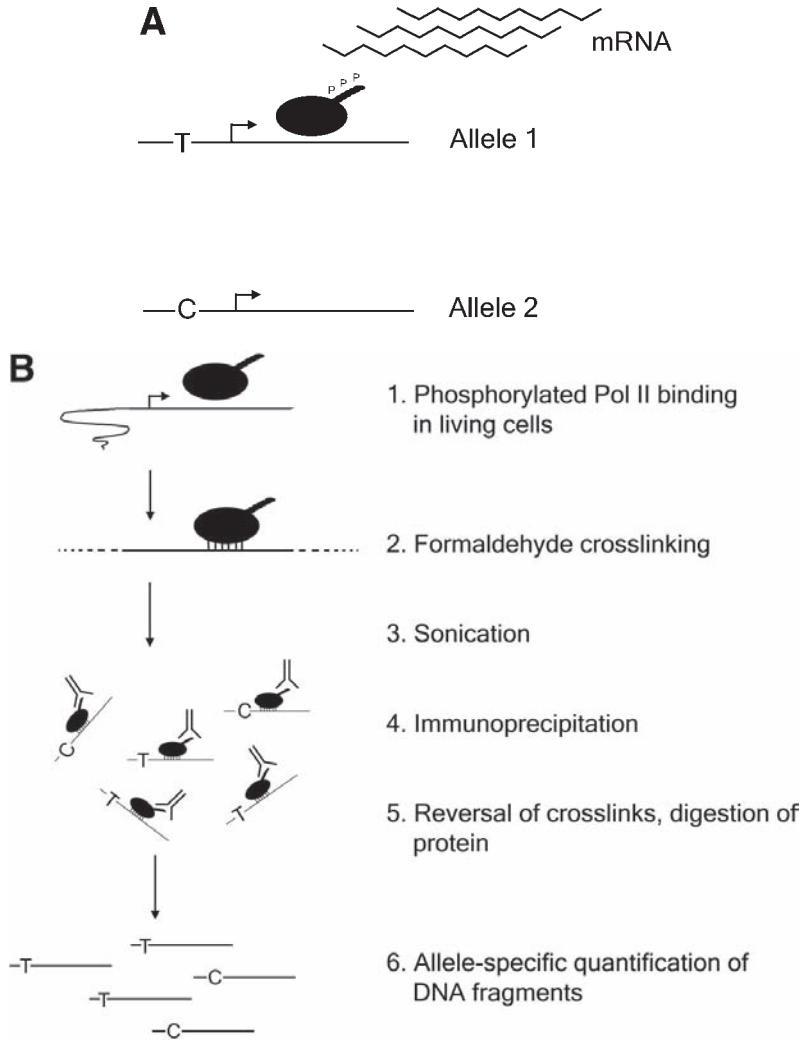


Fig 1. HaploChIP method applied to analysis of phosphorylated Pol II loading. **(A)** In this example, cells originating from an individual heterozygous for a SNP show allele-specific differences in transcriptional activity reflected in mRNA synthesis and phosphorylated Pol II loading. The SNP is used as a marker to distinguish between alleles and may or may not itself be functionally important. Pol II is denoted by the filled oval motif, shown here bound to allele 1. The two allelic forms of the SNP are shown as, T and C, respectively on the two alleles. **(B)** Flow diagram showing steps involved in haploChIP method for this application. Pol II is shown bound to DNA with the two alleles differentiated by the T to C SNP. Immunoprecipitation with antibody specific to the phosphorylated CTD is shown prior to reversal of crosslinks and digestion of protein leaving the DNA fragments, which can be discriminated by the SNP.

ethylpiperazine-*N*'-2-ethanesulfonate, pH 8.0 (Sigma). This buffer can be prepared, autoclaved, and stored at room temperature.
2. 11% Formaldehyde (10X; Merck, Poole Dorset, UK) to be added to 10X formaldehyde crosslinking buffer immediately before use of buffer in a fume hood at room temperature.
3. Glycine (Sigma) solution (20X) prepared to 2.5 *M* final concentration, autoclaved, and stored at room temperature.
4. Phosphate-buffered saline (PBS; Sigma; 1X) autoclaved and stored at 4°C.
5. Freezing buffer: 10% FBS in PBS stored at 4°C.

2.1.3. Isolation of Nuclei

1. Buffer 1 (lysis buffer) (1X): 50 m*M* Hepes (Sigma), 140 m*M* NaCl, 1 m*M* EDTA, 10% glycerol (Sigma), 0.5% NP-40 (Calbiochem, San Diego CA), 0.25% Triton X-100 (Sigma). Store at 4°C.
2. Buffer 2 (1X): 200 m*M* NaCl, 1 m*M* EDTA, pH 8.0, 0.5 m*M* EGTA, pH 8.0, 10 m*M* Tris-HCl, pH 8.0. Store at room temperature.
3. Buffer 3 (1X): 1 m*M* EDTA pH 8.0, 0.5 m*M* EGTA pH 8.0, 10 m*M* Tris-HCl, pH 8.0. Store at 4°C.
4. Complete protease inhibitor tablets (Roche, Mannheim Germany) solubilized in water to 1X immediately before use. Individual protease inhibitors prepared as stock solutions: benzamidine (Sigma) 0.1 *M* in water stored at –20°C and used in buffers 1, 2, and 3 at 1 m*M* final concentration; TLCK (Roche) 1 mg/mL in 0.05 *M* sodium acetate pH 5.0 (Sigma) stored at –20°C and used in buffers 1, 2, and 3 at 50 µg/mL final concentration; TPCK (Roche) 3 mg/mL in ethanol stored at –20°C and used in buffers 1, 2, and 3 at 50 µg/mL final concentration; pepstatin (Roche) 1 mg/mL in ethanol stored at –20°C and used in buffers 1, 2, and 3 at 1 µg/mL.

2.1.4. Sonication

1. Branson 450 Sonifier (Branson Ultrasonics, Branson CT).
2. Prepare 3% *N*-laurosarcosine (Sigma) stock solution.

2.1.5. Purification of Chromatin

1. Prepare cesium chloride (Roche) solutions in TE (10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA): for 1.75 g/cm^3, use 30.294 g of caesium chloride in 30 mL of TE; for 1.5 g/cm^3, use 20.208 g of cesium chloride in 30 mL of TE; for 1.3 g/cm^3, use 12.072 g of cesium chloride in 30 mL of TE. These stock solutions can be stored at room temperature.
2. Use ultraclear centrifuge tubes (Beckman, High Wycombe, Buckinghamshire, UK) for gradient preparation and sample centrifugation
3. Dialysis buffer: 10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA, pH 8.0, 0.5 m*M* of EGTA, pH 8.0, 10% glycerol (Mallinckrodt, Phillipsburg, NJ) and store at 4°C.
4. Spectra/Por dialysis membrane (molecular weight cut off 10K) stored wet in 0.1% sodium azide (Spectrum labs distributed by Fisher Scientific Loughborough Leicestershire, UK). Wash membrane in distilled water then allow to equilibrate

in dialysis buffer for 5 min before loading sample. Ensure there is space between the sample liquid and the clip to allow for swelling during dialysis.

2.2. Immunoprecipitation of Chromatin

1. Antibodies vs phosphorylated serine residues of the carboxy terminal domain of Pol II (Ser5, MMS-134R clone H14; Ser2, MMS-129R clone H5; Covance, Princeton, NJ), anti-T antigen pAb101 (sc-147, Santa Cruz Biotechnology, Santa Cruz, CA).
2. Dynabeads M-280 (Dynal Biotech, Oslo, Norway) precoated with secondary antibody of choice. For example, for antibodies raised in mice, sheep anti-mouse IgG would be convenient and are commercially available (Dynal). Otherwise Dynabeads can be prepared using an antibody of choice in the laboratory following the manufacturer's protocol (Dynal).
3. Magnetic particle concentrator (MPC) suitable for 1.5-mL Eppendorf tubes (Dynal).
4. Prepare a solution of bovine serum albumin (BSA; Sigma) 5 mg per milliliter in PBS (PBS/BSA) immediately before use to wash Dynabeads.
5. Nutator (Shelton Scientific, Shelton, CT).
6. 2X RIPA-POL buffer: 20 m*M* Tris HCl, pH 8.0, 2 m*M* EDTA, 1 m*M* EGTA, 2% Triton X-100 (Sigma), 0.2% sodium deoxycholate (Sigma), 0.2% sodium dodecyl sulfate (SDS; Sigma), 280 m*M* NaCl, 2X complete protease inhibitor, and 10 µg/mL pepstatin. Where indicated, add 100 µg/mL of sonicated herring sperm DNA (Promega, Madison, WI) and 250 m*M* lithium chloride (Sigma).
7. Elution buffer: 10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA, and 1% SDS stored at room temperature.
8. 3 *M* sodium acetate pH 5.2 (Sigma) stored at 4°C.
9. Molecular biology-grade RNase A 10 mg/mL (Roche).
10. Qiagen DNA clean-up kit (Qiagen, Valencia, CA).

2.3. Allele-Specific Quantification

1. For PCR amplification: BioTaq DNA polymerase (Bioline, London, UK), DNA Engine thermal cycler (MJ Research, Waltham, MA).
2. Quantification of relative allelic abundance by primer extension using the massEXTEND platform (Sequenom, San Diego, CA) with detection by MALDI-TOF MS. This includes SpectroCLEAN (Sequenom) resin, SpectroCHIP (Sequenom) microarray, SpectroPOINT (Sequenom) nanoliter dispenser, and a SpectroREADER (Sequenom) mass spectrometer.

3. Methods

3.1. Preparation of Crosslinked Chromatin

3.1.1. Cell Culture

1. This method describes the protocol for lymphoblastoid cell lines grown in suspension but may be applied to other cell types (*see* **Note 1**). Lymphoblastoid cell

lines are grown in suspension to 5×10^8 cells per time point (*see* **Note 2**) and may be stimulated according to experimental design for appropriate gene induction, for example with phorbol 12-myristate 13-acetate and ionomycin.
2. Cells are harvested in mid-log phase. Culture flasks should be taken directly from tissue culture incubator and placed in a fume hood prior to addition of formaldehyde buffer. If needed, an aliquot of cells in suspension may be taken for subsequent RNA harvesting at this point.

3.1.2. Formaldehyde Crosslinking

1. Formaldehyde crosslinking buffer (10X) containing fresh formaldehyde is added directly to tissue culture flasks containing cells in suspension at room temperature in a fume hood (*see* **Note 3**). The flasks should be gently inverted to mix the buffer with the tissue culture medium containing the cells. Allow the mixture to incubate for 45 min at room temperature, inverting flasks at 15 and 30 min (*see* **Note 4**).
2. Add 2.5 *M* glycine (20X) to a final concentration of 125 m*M* directly to cells in formaldehyde buffer to quench the crosslinking reaction. The medium will turn yellow immediately due to the pH change.
3. Collect cells by centrifugation at 500*g* for 5 min at 4°C. Discard supernatant and resuspend cell pellet in 50 mL of cold PBS and repeat centrifugation step. Resuspend in PBS with 10% fetal calf serum (*see* **Note 5**), repeat centrifugation and as much supernatant as possible. Store cell pellets at –80°C.

3.1.3. Isolation of Nuclei

1. Chill all buffers except buffer 2 and keep on ice. Add 1X complete protease inhibitor, benzamidine, TLCK, TPCK, and pepstatin to buffers immediately before use (*see* **Note 6**).
2. Thaw cells quickly in a beaker with chilled water. Resuspend cells thoroughly on ice in 20 mL buffer 1 (lysis buffer) containing protease inhibitors per 5×10^8 cells. Rock at 4°C for 10 min.
3. Pellet in tabletop centrifuge at 2300*g* for 10 min at 4°C. Resuspend pellet in 16 mL of buffer 2 (containing protease inhibitors) at room temperature. Rock gently at room temperature for 10 min.
4. Pellet in tabletop centrifuge at 2300*g* for 10 min at 4°C. Resuspend nuclear material on ice in 4 mL of buffer 3 (containing protease inhibitors).

3.1.4. Sonication

1. Sonicate the suspension of nuclear material in 4 ml aliquots in a 15-mL conical tube, on ice. Use a microtip attached to a Branson 450 Sonifier starting at setting 4 (six times) then increase to 5 (six times), constant power with a 30-s constant burst, allowing the suspension to cool on ice for 1 min between pulses (*see* **Note 7**).
2. This sonication should result in DNA fragments of approx 500 bp to 3 kb in size. To assess this, run 10 to 12 µL of sonicated material on a 1.4% TAE agarose gel. To avoid the sample aggregating in the wells of the gel, add sarkosyl to 0.5% (final concentration) to each sample before loading on the gel.

3. If the average size of chromatin fragments is greater than several kilobytes, repeat sonication as described previously. After dialysis and reversal of crosslinks, the average fragment size is generally 0.5 to 1 kb.

3.1.5 Purification of Chromatin

1. Adjust the suspension to 0.5% sarkosyl by adding 0.02 g of dry sarkosyl powder to 4 mL of suspension, mix well until goes into solution, then rock for 10 min at room temperature.
2. Spin out the debris at 12,000*g* for 10 min at 4ºC and transfer immediately into a new tube.
3. Purify the chromatin by CsCl centrifugation (*see* **Note 8**). CsCl gradients are formed by gently overlaying successively less dense solutions in an ultraclear centrifuge tube (*see* **Note 9**). A total of 4.5 mL of 1.75 g/cm^3 cesium chloride solution is placed at the base of the tube, overlaid with 2.3 mL of 1.5 g/cm^3 cesium chloride solution, which is in turn overlaid with 1.5 mL of 1.3 g/cm^3 cesium chloride solution. The gradient should be used within 30 min of preparation. Proceed to gently overlay 3 mL of chromatin sample onto the top of each cesium chloride gradient. Centrifuge chromatin in an SW41 rotor at 55,000*g* at 20°C for 24 h (*see* **Note 10**).
4. After centrifugation, collect 1-mL fractions by gently drawing off fractions from the top of the gradient with a P1000 tip. It may be necessary to remove the dense suspension near the top of the gradient using a serological pipet or syringe prior to collection of gradient fractions. The top fraction is number 1. Protein and nucleic acid will distribute throughout the gradient. Chromatin usually sediments between fractions 5 and 7. The final fractions will contain small DNA fragments that have not been crosslinked as well as small RNA fragments. Run 12 µL of every other fraction on a 1% agarose TAE gel to check the position of the chromatin (*see* **Note 11**).
5. Pool those samples containing chromatin (generally a total of 3 fractions) and dialyze overnight at 4°C using dialysis buffer (10 m*M* Tris-HCl, pH 8.0, 1 m*M* EDTA, 0.5 m*M* EGTA, 10% glycerol). Typically, three samples of chromatin can be dialyzed in 1 liter of dialysis buffer. The next morning, change the buffer once and continue dialysis against 1 liter of fresh buffer for an additional 4 h. Conductivity can be checked to ensure that dialysis is complete.
6. Transfer chromatin from dialysis membrane into polypropylene round-bottomed tube. Spin at 12,000*g* for 10 min in SS-34 rotor. Transfer chromatin into a new tube and divide into 300-µL aliquots (each of which should contain 300–500 µg) in Eppendorf tubes, snap freeze in liquid nitrogen and store at –80°C or use immediately for immunoprecipitation.

3.2. Immunoprecipitation of Chromatin

3.2.1. Preparation of Magnetic Bead-Antibody Complex and Immunoprecipitation

1. Typically, a standard individual immunoprecipitation reaction would use 50 µg of cesium–chloride-purified chromatin with 2 µg of primary antibody (such as vs

phosphorylated Pol II) attached to 50 µL of starting volume of Dynabeads. The experimental design should include appropriate negative controls.
2. Take sufficient Dynabeads M280 precoated with secondary antibody for the number of immunoprecipitation reactions proposed in a 1.5-mL Eppendorf tube and concentrate in MPC by placing in MPC rack for 2 min, tilting rack to horizontal at 1 min to remove any beads trapped in lid. Remove supernatant with a pipet and resuspend Dynabeads in 1 mL of PBS/BSA. Repeat, washing a total of two times.
3. Add sufficient primary antibody in 1 mL of PBS/BSA to the Eppendorf tube and incubate overnight on a nutator (setting 4) at 4°C.
4. To remove any unbound antibody, place tube in MPC, concentrate for 2 min and remove supernatant as previously. Resuspend antibody-bound beads in 1 mL of PBS/BSA and repeat wash step a total of three times using MPC.
5. Resuspend in volume of PBS/BSA equal to starting volume of Dynabeads taken from stock.
6. Set up immunoprecipitation by combining chromatin with 2X RIPA-POL buffer, TE and Dynabeads bound to primary antibody so as to make final concentration 1X RIPA-POL in a total volume of 500 µL.
7. Incubate overnight on a nutator (setting 4) at 4°C.

3.2.2. Washing Immunoprecipitations

1. Place Eppendorf tubes containing chromatin with antibody bound to beads in MPC for 2 min to concentrate the beads (*see* **Note 12**).
2. Save 50 µL of chromatin (supernatant above concentrated beads when in MPC) from antibody negative control tube and transfer to a new Eppendorf. This sample will be valuable later to estimate final chromatin fragment size. Add 50 µL of TE and put on ice.
3. Save 2.5 µL of chromatin from antibody negative control tube and transfer to a new Eppendorf. This sample will be used later as input control. Add 97.5 µL of TE and put on ice.
4. Aspirate unbound chromatin using a Pasteur pipette on a vacuum. Wash bead complexes with 1 mL of freshly prepared 1X RIPA-POL buffer. Add buffer to each tube and invert tubes twice to resuspend.
5. Use MPC to precipitate the beads at room temperature. Repeat 1X RIPA-POL wash step.
6. Wash precipitates with 1 mL of freshly prepared 1X RIPA-POL buffer with 100 µg/mL herring sperm DNA and rock at room temperature for exactly 5 min.
7. Wash precipitates with 1 mL of freshly prepared 1X RIPA-POL buffer with 100 µg/mL herring sperm DNA plus 300 m*M* NaCl and rock at room temperature for exactly 5 min.
8. Wash precipitates with 1 mL of 1X RIPA-Pol with 250 m*M* LiCl and remove all traces of wash solution.
9. Wash once with 1 mL of TE in MPC then gently pellet beads by centrifugation to remove any remaining liquid with a pipet.

3.2.3. Elution From Beads and Reversal of Crosslinks

1. Add 50 µL of elution buffer to beads, vortex briefly to resuspend, and incubate at 65°C for 10 min. Vortex briefly every 2 min during incubation.
2. Spin for 30 s at maximum speed in a Microfuge and transfer supernatant to a new Eppendorf tube. Discard the bead pellet.
3. Add 120 µL of elution buffer to the supernatant in the new tube. Reverse crosslinks at 65°C overnight in water bath.
4. For input controls, add 11 µL of 10% SDS and reverse crosslinks at 65°C overnight in water bath.

3.2.4. DNA Extraction and Precipitation

1. Add 150 µL of TE containing proteinase K (final concentration 200 µg per mL) to each Eppendorf tube and incubate for 2 h at 37°C
2. Extract with an equal volume of equilibrated phenol-chloroform pH 8.0, vortex and spin at maximum speed in a Microfuge for 5 min. Repeat phenol–chloroform extraction.
3. Extract once with an equal volume of chloroform–isoamyl alcohol.
4. Precipitate DNA fragments by addition of one-tenth the volume of 3 *M* sodium acetate, pH 5.2, 2.5X volume of ice-cold 100% ethanol, 20 µg glycogen, and vortex briefly. Incubate at –20°C overnight. Spin at maximum speed in a Microfuge for 10 min at 4°C then wash pellet with 1 mL of ice-cold 70% ethanol, vortex, spin for 5 min at 4°C at maximum speed. Air dry pellet.
6. Resuspend pellet in 30 µL of TE containing 10 µg of RNase A and incubate for 1 h at 37°C.
7. Remove RNase by spin column purification according to manufacturer's instructions and elute in 10 m*M* Tris pH 8.0. Store material at –20°C.

3.3. Allele-Specific Quantification

1. The products of chromatin immunoprecipitation should first be assayed in a nonallele-specific manner using primers specific for the gene of interest by semiquantitative PCR ***(10)***, multiplexing with a housekeeping gene where appropriate.
2. Allele-specific quantification can be achieved using a number of different approaches of which primer extension with detection by mass spectrometry will be described here (*see* **Note 13**). A number of steps are involved which potentially introduce variance into the experiment: for this reason duplication should be included at each step (*see* **Note 14** ***[11,12]***).
3. For primer extension with detection by mass spectrometry, a first round PCR is performed using 5 ng of genomic DNA or 5 µL of chromatin immunoprecipitation (ChIP) DNA in a 25-µL reaction volume using 0.5 U of BioTaq with 0.8 m*M* dNTPs, 1.9 m*M* $MgCl_2$, and 0.2 µ*M* each primer. Thermal cycling parameters should be optimized to ensure cycle number remains in the linear phase of amplification with annealing temperatures dependent on the primer design: for an MJ

Tetrad a typical cycling protocol would be 96°C for 1 min followed by 6 cycles of 94°C for 45 s, 56°C for 45 s, 72°C for 30 s; then 30 cycles of 94°C for 45 s, 65°C for 45 s, 72°C for 30 s; followed by final extension at 72°C for 10 min.

4. The PCR product is then subaliquoted onto a 384-well plate, and nonincorporated dNTPs removed using shrimp alkaline phosphatase by incubating at 37°C for 20 min followed by 85°C for 5 min. Primer extension is performed using a homogeneous MassEXTEND (Sequenom) reaction comprising a cocktail of 100 μ*M* extension primer, 0.576 U MassEXTEND enzyme, buffer and an appropriate deoxy and dideoxy nucleotide termination mix. A typical primer extension reaction would comprise 94°C for 2 min then 40 cycles of 94°C for 5 s, 52°C for 5 s, 72°C for 5 s.
5. The products of primer extension are desalted using SpectroCLEAN (Sequenom) resin and transferred onto a SpectroCHIP (Sequenom) microarray by SpectroPOINT (Sequenom) nanoliter dispenser. MALDI-TOF analysis is performed using a SpectroREADER (Sequenom) mass spectrometer.

4. Notes

1. Other cell types growing in suspension can be harvested as described. For adherent cell lines, formaldehyde crosslinking buffer (10X) containing formaldehyde should be added directly to tissue culture dishes containing media, incubated and quenched with glycine (20X) as noted. Cells can then be washed using PBS *in situ* by aspiration of media/buffer and detached using a cell scraper.
2. The number of cells required to generate sufficient chromatin for successful immunoprecipitation and PCR is 5×10^8 cells in our hands; it may prove possible to successfully apply the approach to 1×10^8 cells. Given the scale of tissue culture, investigators may find use of roller bottle cultures helpful.
3. It is vital that fresh formaldehyde be used from an unopened stock. This is a hazardous chemical and care should be used in its storage, use and disposal with handling only in a fume hood.
4. The time and temperature of crosslinking with formaldehyde should be optimized for a given cell line and type. For example, between 5 and 60 min at either room temperature or on ice.
5. The use of FBS in the final wash is optional but helps to protect chromatin integrity in freeze-thawing.
6. For haploChIP experiments analyzing phosphorylated Pol II, 10 m*M* (final concentration) sodium pyrophosphate (Sigma) should be used in all buffers including buffers 1–3, and those used subsequently in dialysis and for immunoprecipitation.
7. To minimize foaming during sonication, place probe at mid-depth in suspension then turn on the power; start at lower settings then increase.
8. It may be possible to omit the cesium chloride purification step for certain applications. This appears to be dependent on the quality and affinity of the antibody used in the immunoprecipitation step and the nature of the protein to be immunoprecipitated. For example, histones are very amenable to chromatin immunoprecipitation and immunoprecipitate well without the need for a cesium chloride

purification step. Instead, to prepare chromatin for immunoprecipitation of proteins that immunoprecipitate robustly, the sonicated material can simply be centrifuged at 12,000*g* at 4°C for 10 min and adjusted to 10% final glycerol concentration prior to storage at –80°C and use directly in immunoprecipitation reactions.

9. A cesium chloride gradient can be generated by using a 3-mL syringe with a 18-gage needle and very slowly preparing the gradient by layering successive solutions. Avoid drops and keep the needle tip just below the surface.
10. The success of cesium chloride purification is dependent on a number of factors including: 1) careful preparation of the cesium chloride gradient; 2) ensuring a thin coat of lubricant is applied to the centrifuge bucket threads; 3) removing the bucket gaskets (O rings) and coating them lightly but evenly with silicone vacuum grease; 4) drying the exterior of the tubes before placing in the centrifuge buckets as moisture between the bucket and tube can cause the tube to collapse; 5) pairs of ultracentrifuge tubes must be equal in weight and filled to within 3 mm of the top of the tube.
11. The chromatin will run anomalously high because of the high concentrations of CsCl and protein. Chromatin should run at approximately the middle of the gradient and will appear as a large smear of DNA 1 kb or higher in size (if the sample is de-proteinized and desalted, it will run smaller at approx 500 bp). Protein and detergent will be present near the bottom of the gradient while RNA can be found toward the bottom.
12. The washing of immunoprecipitation reactions is critical to the success of the experiment and should be done as carefully as possible to minimize nonspecific background. The stringency of salt and SDS concentration that can be used may need to be titrated for the individual protein of interest; the protocol given here is a useful starting point and should work well for most transcription factors.
13. The method used for allele-specific quantification will depend on facilities available in the laboratory. A number of approaches for accurate quantification of cDNA have been published for allele-specific quantification, notably using single base extension methods in the presence of fluorescently labeled nucleotides with detection on a DNA sequence detector (***5***,***11***) or by restriction enzyme digestion and real-time PCR amplification (***12***).
14. For PE/MS of a given haplotypic set, independent cell lines should be studied and replicate immunoprecipitations performed. A given immunoprecipitated sample of DNA fragments should then be analyzed by independent PCR, which are in turn spotted as replicates on the 384 well detection chip and analyzed by the mass spectrometer with independent reads.

Acknowledgments

I am grateful to T. Maniatis, B. Dynlacht, J. Rayman, T. Kim, and B. Ren for their advice in establishing this protocol. This work was supported by the Medical Research Council UK.

References

1. Knight, J. C., Keating, B. J., Rockett, K. A., and Kwiatkowski, D. P. (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat. Genet.* **33,** 469–475.
2. Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002) Finding genes that underlie complex traits. *Science* **298,** 2345–2349.
3. Cargill, M., Altshuler, D., Ireland, J., et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22,** 231–238.
4. Knight, J. C. (2003) Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. *Clin. Sci. (Lond)* **104,** 493–501.
5. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. (2002) Allelic variation in human gene expression. *Science* **297,** 1143.
6. Knight, J. C., Keating, B. J., and Kwiatkowski, D. P. (2004) Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1. *Nat. Genet.* **36,** 394–399.
7. Orlando, V., Strutt, H., and Paro, R. (1997) Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* **11,** 205–214.
8. Jurinke, C., van den Boom, D., Cantor, C. R., and Koster, H. (2002) Automated genotyping using the DNA MassArray technology. *Methods Mol. Biol.* **187,** 179–192.
9. Braun, A., Little, D. P., and Koster, H. (1997) Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin. Chem.* **43,** 1151–1158.
10. Takahashi, Y., Rayman, J. B., and Dynlacht, B. D. (2000) Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression. *Genes Dev.* **14,** 804–816.
11. Cowles, C. R., Joel, N. H., Altshuler, D., and Lander, E. S. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.* **32,** 432–437.
12. Weber, M., Hagege, H., Lutfalla, G., et al. (2003) A real-time polymerase chain reaction assay for quantification of allele ratios and correction of amplification bias. *Anal. Biochem.* **320,** 252–258.

III

Genotyping Techniques

6

Aspects Influencing Genotyping Method Selection

Peter Imle

Summary

The variety of genotyping methods currently available and the evolution of their capabilities have facilitated an expansion of the field of pharmacogenomics. Traditionally, limited genotyping capabilities have restricted the generation and application of genotyping data for pharmacogenomic studies. With the variety of platforms and chemistries available for flexible, high-throughput genotyping, it is important to keep in mind the limitations imposed by both the polymorphisms that are to be interrogated and the type of pharmacogenomics study for which the data are being generated. This chapter is an overview of the constraints these factors impose on different genotyping methods and describes aspects important to the integration of genotyping into a pharmacogenomics study.

Key Words: Genotyping; polymorphism; single nucleotide polymorphism (SNP); simple sequence repeat (SSR); tandem repeat; insertion; deletion.

1. Introduction

The completion of the human genome draft sequence and the ongoing annotation of genes and genetic variants have provided the foundation for an expanding role of genotyping in pharmacogenomics. Detailed information on genomic variants of all types is being constantly compiled and should be considered an invaluable resource for the identification and screening of these polymorphisms and their association with clinical phenotypes ***(1–7)***. As this wealth of genomic information has grown, so have the number of methods available for genomic interrogation. The effective integration of genotyping data into a pharmacogenomics study requires an understanding of the factors influencing the efficiency of different genotyping methods and the priorities required of different study designs. A variety of genotyping chemistries are available for interrogating all types of deoxyribonucleic acid (DNA) sequence variants ***(8–10)***. The physical nature of the variants in question will render some techniques uninformative. Because of this, the selection of a genotyping

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

chemistry requires an initial assessment of the types of variants that are going to be investigated. The type of pharmacogenomics study for which the genotyping data are being generated also will limit the usefulness of some of the available methods. Different study designs have different priorities for the number of assays necessary, the stringency of their physical location, and their minimum call rate. Prioritizing these aspects of the investigation at the outset can greatly increase the effectiveness of the genotyping method selected and in turn efficiently generate more useful data.

The first portion of this chapter is dedicated to a discussion of different types of genomic variants and their effects on selection of a genotyping technique. The later sections highlight the different types of pharmacogenomic studies and the unique constraints they apply to selection of a genotyping technique. The aspects influencing method selection covered in this chapter are considered independent of platform as the application of different methods are covered in-depth in the following chapters. It also should be mentioned that often the overriding influence on the application of a genotyping method is dictated by the platforms that are available rather than its suitability for a given genotyping project. In this instance it is still important to keep the following information in mind in order to maximize the application of the available technology to new investigations in Pharmacogenomics.

2. Types of Variation

Genome databases currently contain information on millions of genetic variants located throughout the genome ***(1–7)***. These variants can be classified both by their effect on nearby genes and by their physical attributes. It is these classifications that dictate the application of different genotyping chemistries. The investigation of single nucleotide polymorphisms (SNPs) often requires a different chemistry than more substantial DNA variations, such as tandem repeats. Careful consideration of these aspects in the experimental design process will aid in the generation of high quality genotyping data.

2.1. Functional Classification

Variants fall into two categories, functional and nonfunctional, based whether the change at the genomic sequence level has an effect on normal gene function ***(11)***. The magnitude of the variant at the sequence level can range from single base changes to the deletion of entire genes. The resulting functional effect can be equally as diverse and manifest as changes in the amino acid sequence, vary rates of transcription by modification of promoter sequence, vary the sequence of intron/exon boundaries resulting in alternatively spliced ribonucleic acid (RNA) products, or cause the deletion of entire genes ***(11,12)***. These functional changes represent the genetic portion of observed clinical

phenotypes ***(12)***. Thus, the identification and characterization of the functional variants responsible for the clinical phenotypes involved in a pharmacogenomics study are the ultimate goal.

Although they do not directly influence the clinical phenotype, nonfunctional variants can be critical to the effective application of genotyping to a pharmacogenomics study. Nonfunctional variants are much more abundant genome wide as there is no genetic selection against their accumulation in the population ***(13,14)***. Although the individual alleles of a given nonfunctional variant may have no direct impact on normal function of a particular gene, its physical location may place it in linkage disequilibrium with a functional variant. This provides the opportunity to use the more common nonfunctional variants as markers to localize novel functional variants via linkage analysis and association studies. Additionally, their relative abundance can provide the opportunity to be used in place of directly screening the functional variants for which validating a high performance genotyping assay cannot be achieved.

2.2. Physical Classification

2.2.1. Single Nucleotide Polymorphisms

SNPs are the most common type of genetic variant ***(11)***. An estimated 10 million or more SNPs are dispersed throughout the human genome ***(15)***. Their abundance and consistent occurrence throughout the genome have made them an increasingly useful target for pharmacogenomic investigations. More recently their utility has been reinforced by the accumulation of evidence, that suggests much of the heritable portion of common disease is attributed to minor sequence variations as opposed to more substantial gene defects such as insertions or deletions ***(16)***. As a natural recourse to their abundance, there is a dispersal of SNPs of both nonfunctional and functional significance. From the perspective of their applicability to genotyping in pharmacogenomic studies, this provides a critical resource. Currently, there is information on more than 9 million SNPs deposited in public databases ***(2)***. Projects are underway that are compiling genotype, gene frequency and, in some cases, validated assays for a variety of SNPs throughout the genome ***(1–7)***. These data should be considered a primary resource for any project that will involve SNP genotyping.

2.2.2. Simple Sequence Repeats

Simple sequence repeats (SSR) are stretches of DNA that are composed of various numbers of repeated DNA segments in which the core repeat unit is 1 to 4 bp in length. These variants are dispersed throughout the genome, are shown to be both functional and nonfunctional in nature, and are highly variable, resulting in the accumulation of multiple alleles for each SSR locus ***(17)***.

These factors make many of these loci very informative as markers for studies of association and linkage analysis. Commonly referred to as microsatellites, these markers have been used extensively for mapping, association, and human identification studies since their discovery ***(18–20)***. Studies of population stratification have suggested that on average the genotypic information available from a single microsatellite marker is equivalent to that provided by five to eight SNP markers ***(21)***. The primary drawback of using microsatellites for mapping is that they are not as common as SNPs; thus, there may not be the density of markers necessary to investigate a specific region of interest. Additionally, as more data are generated on the number, location, and functional significance of SNPs in the human genome, it has been shown that specific SNPs can be even more informative than some dinucleotide microsatellite markers ***(21)***. Combined with increasing efficiency of genotyping SNPs the advantage of microsatellites being more informative is being overcome.

Specific SSRs have been associated with variation in gene function and clinically observed phenotypes ***(22,23)***. As would be expected, the expansion or contraction of a region of DNA sequence can have functional consequences that are highly dependent on its proximity to coding or regulatory DNA sequence. Different alleles of a 2-bp SSR in the 5' promoter region of the gene UGT1A1 have been shown to be associated with changes in efficiency of irinotecan metabolism ***(22,23)***. A common method of genotyping of this polymorphism involves the measurement of polymerase chain reaction (PCR)-amplified fragments that contain the polymorphic sequence ***(22)***. Two base pair variations in length of the amplified fragment can be detected by capillary electrophoresis or some gel electrophoresis-based systems ***(24)***.

2.2.3. Variable Number Tandem Repeats

Variable number tandem repeats (VNTRs) are the sequential and repeated insertion of DNA sequences whose basic repeat unit is larger than those represented by SSRs. These tandem repetitions of DNA sequence can be tens of bases long and their insertion can rapidly change the structure of a gene at the sequence level. Because of the more substantial alteration of gene sequence caused by the VNTR polymorphisms, genotyping methods generally have focused on the detection of size variation in PCR amplicons containing the repeats ***(25,26)***. Selection of a genotyping method for the screening of these types of polymorphisms must take into account not only the length of the repeat sequence, but also the number of repeats likely to be present in the study population. Because the number of repeat units increases so does the length of the PCR product. Careful consideration must be taken to keep the resulting amplicons within the detection capabilities of the genotyping platform.

The presence of a 28-bp VNTR in the promoter region of the thymidylate synthase (TYMS) gene has been shown to be associated with changes in in

vitro expression levels of the gene ***(26)***. TYMS is an important target for several chemotherapy drugs ***(25)***. Genotyping for the TYMS promoter VNTR has commonly been performed by analysis of variation in PCR amplicon length to calculate the number of repeat units ***(25,26)***. This method is robust; however, special consideration must be taken to keep the length of the amplified fragments within the detectable range of the platform being used for analysis. Thus far, there have been up to nine repeats of the 28-bp repeat identified, which requires the detection of fragments varying throughout a range of 250 bp, not including additional space required for primer location.

2.2.4. Insertions and Deletions

The most profound of the sequence variations described here, large-scale insertions and deletions, can have the most noticeable phenotypic effects. Insertions of DNA sequence within a gene can have the same deleterious effects on gene function that have been discussed earlier, ranging from changes in gene transcription rates to total elimination of gene function. An excellent example of the spontaneous insertion of DNA sequence is represented in the cytosine beta synthase gene, in which associations have been made between the insertion of a 68-bp segment of DNA sequence from a previous intron/exon junction and observed changes in plasma total homocysteine ***(27,28)***. The analysis of PCR fragment size has been commonly used for genotyping of this polymorphism by methods similar to that mentioned previously ***(27–29)***.

Deletion of genomic sequence can have substantial repercussions as well. The elimination of DNA sequence can vary in magnitude from single base pairs to entire genes. Various methods are capable of genotyping these variants; however, the deciding factor on their application lies in the conservation of the sequence surrounding the deletion junction. Pyrosequencing has been used successfully to interrogate deletions as long as 100 bp ***(30)***, and the analysis of changes in amplicon size of PCR products containing the deleted sequence has been used for detecting deletions up to a few hundred base pairs ***(31,32)***. A primary benefit of these two methods for the genotyping of deletions is the capability to detect heterozygosity in samples.

3. Types of Pharmacogenomics Studies

Early studies of pharmacogenomics focused on the assignment of Mendelian inheritance patterns to observed clinical phenotypes ***(33)***. This method was limited to finding mostly monogenic traits that showed simple inheritance patterns identifiable in relatively small treatment groups. Advances in the generation of genomic, functional genomic, and proteomic data have lead to the identification and classification of a wealth of gene ontology and pathway information. These resources have allowed for the identification of candidate genes that are likely to be involved in the observed clinical phenotypes

(33,34). Methodologies for interrogation of these specific loci have made possible the study of more complex, multi-locus genetic interactions. This method of screening candidate genes has been successfully applied to Pharmacogenomic studies of drug metabolizing genes in a variety of diseases *(34–39)*. These efforts have identified both associations between the variants and clinical phenotypes as well as the novel variants responsible for the functional changes.

Some studies in pharmacogenomics have taken a broader approach to the identification of significant genetic variants by scanning the entire human genome *(38)*. The primary limitation for this technique is the need to do a large number of assays to get adequate coverage of the complete genome. Estimates of the number of markers necessary for this type of study are highly dependent on sample size, but can quickly exceed 50,000 markers necessary to have adequate power of detection *(38,40)*. Because many studies have a predefined number of available samples, statistical analysis can be performed to estimate the detection capabilities of various numbers of maker within the study group *(40,41)*.

The implications that each of these methods have on the selection of a genotyping technology is summarized **Table 1**. These classifications are relative to one another and highlight the strengths of these different study designs. The whole genome approach requires a method that can efficiently process the larger number of assays that are required to provide sufficient power of detection and coverage of the entire genome. Although a greater number of assays may be required, the constraints on successful assay validation are relaxed in comparison with the candidate gene approach as a result of the prevalence of nonfunctional variants in the genome. If a robust assay cannot be validated for a specific polymorphism, it is likely that there will be another variant nearby that can be substituted. Additionally, the assays used for the whole genome approach are not required to be as robust because a slight drop in call rate for an individual polymorphism will not have as profound an effect on the data set as it would with the more demanding candidate gene study.

4. Conclusion

The current age of genotyping finds itself growing on the foundation of the human genome draft sequence but not yet at the point of cost effectively obtaining whole genome sequence information on an individual basis. This requires that genotyping methods exist to do both screening of known variants and identification of novel polymorphisms at levels of cost and throughput such that the data can be efficiently generated and effectively used in studies of pharmacogenomics. The applicability of different genotyping methods changes as the demands of the genotyping project are defined. Important aspects for consideration prior to method selection are the type of variants that are to be

Table 1
Implication of Different Types of Pharmacogenomics Studies on Aspects of Assay Design

	Type of Study	
	Candidate gene	Whole genome
Number of polymorphisms	Moderate	High
Assay validation rate	High	Moderate
Individual assay call rate	High	Moderate

These assessments are relative to the two study designs and highlight the comparative strengths of each. The candidate gene approach may require fewer assays in total but has a higher stringency for the performance of the assays. These parameters are relaxed somewhat for whole genome analysis, but at the expense of having to do a larger number of assays in total.

interrogated and the type of pharmacogenomics study that the data will be integrated. Clear definition of these parameters will greatly aid in the selection and appropriate application of a genotyping method. Continuing improvement of genotyping methods, specifically reduction in cost and increases in both throughput and efficiency of template use, are central to continued advancement of genotyping in Pharmacogenomics studies.

The following chapters cover in detail the application of several different genotyping methods currently used in pharmacogenomic studies. These methods use a variety of different chemistries and platforms for the screening of genetic variants, and all have different advantages based on the constraints outlined in this chapter. Understanding of these constraints and their effect on the genotyping capabilities required of the project, combined with the thorough consideration of genotyping in the experimental design process, will greatly aid in the generation of high quality data.

References

1. Kwok, P. Y. and Gu, Z. J. (1999) Single Nucleotide polymorphism libraries: why and how are we building them? *Mol. Med. Today* **5,** 538–543.
2. Wheeler, D. L., Church, D. M., Edgar, R., et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32,** D35–40. Available at: http://www.ncbi.nlm.nih.gov/SNP/.
3. Birney, E., Andrews, T. D., Bevan, P., et al. (2004) An overview of Ensembl. *Genome Res.* **5,** 925–928. Available at: http://www.ensembl.org.
4. The international SNP map working group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933. Available at: http://snp.cshl.org/.
5. The international hapmap consortium. (2003) The international hapmap project. *Nature* **426,** 349–357. Available at: http://www.hapmap.org.

6. Packer, B. R., Yeager, M., and Staats, B., et al. (2004) SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res*. **32,** D528–D32. Available at: http://snp500cancer.nci.nih.gov/home_1.cfm?CFID=2241&CFTOKEN=69829203.
7. Abeysinghe, S. S., Stenson, P. D., Krawczak, M., and Cooper, D. N. (2004) Gross rearrangement breakpoint database (GRaBD). *Hum. Mutat.* **23,** 219–221. Available at: http://www.uwcm.ac.uk/uwcm/mg/grabd/.
8. Kwok, P. Y. (ed.) (2003) *Single Nucleotide Polymorphisms*. Humana, Totowa, NJ.
9. Mitchelson, K. R. and Cheng, J., (ed.) (2001) *Capillary Electrophoresis of Nucleic Acids, Vol II*. Humana, Totowa, NJ.
10. Kwok, P. Y. (2000) High-throughput genotyping assay approaches. *Pharmacogenomics*. **1,** 95–100.
11. Brooks, L. (2003) SNPs: why do we care, in *Single Nucleotide Polymorphisms* (Kwok, P. Y., ed.), Humana, Totowa, NJ, pp. 1–14.
12. Wang, Z. and Moult, J., (2001) SNPs, protein structure, and disease. *Hum. Mutat*. **17,** 263–270.
13. Cargill, M., Sltchuler, D., Ireland, J., et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet*. **22,** 231–238.
14. Chakravarti, A. (1999) Population genetics—making sense out of sequence. *Nat. Genet*. **21,** 239–247.
15. Kruglyak, L. and Nickerson, DA. (2001) Variation is the spice of life. *Nat. Genet*. **27,** 234–236.
16. Imyanitov, E. N., Togo, A. V., and Hanson, K. P. (2004) Searching for cancer-associated gene polymorphisms: promises and obstacles. *Cancer Lett*. **2004,** 3–14.
17. Schlotterer, C. (2004) The evolution of molecular markers—just a matter of fashion? *Nat. Rev. Genet*. **5,** 63–69.
18. Jonasdottir, A., Thorlacius, T., Fossdal, R., et al. (2003) A whole genome association study in Icelandic multiple sclerosis patients with 4804 markers. *J. Neuroimmunol*. **143,** 88–92.
19. Collins, J. R., Stephens, R. M., Gold, B., Long, B., Dean, M., and Burt, S. K. (2003) An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*. **82,** 10–19.
20. Nievergelt, C. M., Smith, D. W., Kohlenberg, J. B., and Schork, N. J. (2004) Large-scale integration of human genetic and physical maps. *Genome Res*. **14,** 1199–1205.
21. Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003) Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet*. **73,** 1402–1422.
22. Monaghan, G., Ryan, M., Seddon, R., Hume, R., and Burchell, B. (1996) Genetic variation in bilirubin UDP-glucuronocyltransferase gene promoter and Gilbert's syndrome. *Lancet* **347,** 578–581.
23. Bosma, P. J., Chowdhury, J. R., Bakker, C., et al. (1995) The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N. Engl. J. Med.* **333,** 1171–1175.
24. Mansfield, E. S., Wilson, R. B., and Fortina, P. (2001) Analysis of short tandem repeat markers by capillary array electrophoresis, in *Capillary Electrophoresis of*

Nucleic Acids, Vol II. (Mitchelson, K. R., and Cheng, J., eds.), Humana, Totowa, NJ, pp. 151–162.

25. Etienne, M. C., Ilc, K., Formento, J. L., et al. (2004) Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphisms: relationships with 5-fluorouracil sensitivity. *Br. J. Cancer* **90,** 526–534.
26. Marsh, S., Collie-Duguid, E. S.R., Li, T., Liu, X., and McLeod, H. L. (1999) Ethnic variation in the thymidylate synthase enhancer region polymorphism among Caucasian and Asian populations. Genomics **58,** 310–312.
27. Tsai, M. Y., Bignell, M., Schwichtenberg, K., and Hanson, N. Q. (1996) High prevalence of a mutation in the Cystathionine β-synthase gene. *Am. J. Hum. Genet.* **59,** 1262–1267.
28. Tsai, M. Y., Bignell, M., Yang, F., Welge, B. G., Graham, K. J., and Hanson, N. Q. (2000) Polygenic influence on plasma homocysteine: association of two prevalent mutations, the 844ins68 of cystathionine β-synthase and A2756G of methionine synthase, with lowered plasma levels. *Atherosclerosis* **149,** 131–137.
29. Kraus, J. P., Loiveriusova, J., Sokolova, J., et al. (1998) The human cystathionine ß-sythase (CBS) gene: complete sequence, alternative splicing, and polymorphisms. *Genomics* **52,** 312–314.
30. Guo, D. C., Qi, Y., He, R., Gupta, P., and Milewicz, D. M. (2003) High throughput detection of small genomic insertions or deletions by Pyrosequencing. *Biotechnol. Lett.* **25,** 1703–1707.
31. Huang, X. H., Salomake, A., Malin, R., Koivula, T., Jokela, H., and Lehtimaki, T. (1997) Rapid identification of angiotensin-converting enzyme genotypes by capillary electrophoresis. **43,** 2195–2196.
32. Wenz, H. M., Dailey, D., and Johnson, M. D. (2001) Development of a high-throughput capillary electrophoresis protocol for DNA fragment analysis, in *Capillary Electrophoresis of Nucleic Acids, Vol II*. (Mitchelson, K. R., Cheng, J., eds.), Humana, Totowa, NJ, pp. 3–18.
33. Goldstein, D. B., Tate, S. K., and Sisodiya, S. M. (2003) Pharmacogenetics goes genomic. *Nat. Rev. Genet.* **4,** 937–947.
34. Evans, W. E. and McLoed, H. L. (2003) Pharmacogenomics-drug disposition, drug targets, and side effects. *N. Engl. J. Med.* **346,** 538–549.
35. Evans, W. E. and Relling, M. V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286,** 487–491.
36. Ulrich, C. M., Robien, K., and McLoed, H. L. (2003) Cancer pharmacogenetics: polymorphisms, pathways and beyond. *Nat. Rev. Cancer* **3,** 912–920.
37. Johnson, J. A. (2003) Pharmacogenetics: potential for individualized drug therapy through genetics. *Trends Genet.* **19,** 660–666.
38. McLeod, H. L. and Evans, W. E. (2001) Phamacogenomics: unlocking the human genome for better drug therapy. *Annu Rev. Pharmacol. Toxicol.* **41,** 101–121.
39. Daly, A. K. (2003) Pharmacogenetics of the major polymorphic metabolizing enzymes. *Fundam. Clin. Pharmacol.* **17,** 27–41.
40. Hao, K., Xu, X., Laird, N., Wang, X., and Xu, X. (2003) Power estimation of multiple SNP association test of case-control study and application. *Genet. Epidemiol.* **26,** 22–30.

41. Service, S. K., Sandkuijl, L. A., and Freimer, N. B. (2003) Cost-effective designs for linkage disequilibrium mapping of complex traits. *Am. J. Hum. Genet.* **72,** 1213–1220.

7

Denaturing High-Performance Liquid Chromatography for Mutation Detection and Genotyping

Donna Lee Fackenthal, Pei Xian Chen, and Soma Das

Summary

Denaturing high-performance liquid chromatography (DHPLC) is an accurate and efficient screening technique used for detecting deoxyribonucleic acid sequence changes by heteroduplex analysis. It can also be used for genotyping of single-nucleotide polymorphisms. The high-sensitivity of DHPLC has made this technique one of the most reliable approaches to mutation analysis and is used in various areas of genetics, both in the research and clinical arena. This chapter describes the methods used for mutation detection analysis and the genotyping of single-nucleotide polymorphisms by DHPLC on the WAVE™ system from Transgenomic Inc.

Key Words: Denaturing high-performance liquid chromatography (DHPLC); genotyping; mutation detection; single-nucleotide polymorphism (SNP); single-base extension (SBE).

1. Introduction

1.1. Mutation Detection by DHPLC

The basis of mutation detection by denaturing high-performance liquid chromatography (DHPLC) is the formation and discrimination of homoduplex and heteroduplex deoxyribonucleic acid (DNA) molecules that can be created when a DNA sequence change is present on one allele ***(1)***. The DHPLC column (DNASep®) contains a nonporous matrix consisting of polystyrene–divinylbenzene copolymer beads. The beads are alkylated with C-18 chains that form single C-C bonds, are electrostatically neutral, and do not interact with nucleic acids ***(2)***. DNA binds to the column by the use of triethylammonium acetate (TEAA), which serves as an ion-pairing reagent between nucleic acids and the beads on the column. The positively charged triethylammonium ion

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

bonds to the negatively charged phosphate group on the DNA backbone and the hydrophobic groups of TEAA interact with the hydrophobic C-18 chains on the copolymer beads. DNA is eluted off the column by the use of an acetonitrile buffer, which at increasing concentrations across the column breaks the hydrophobic interactions between the TEAA–DNA molecules. Because heteroduplex molecules form less hydrophobic interactions compared with homoduplex molecules, they are eluted off the column faster compared with homoduplex molecules.

Coding exons and flanking intron sequences generally are targeted for mutation detection. These regions are amplified with specific primers to create amplicons of approx 180–700 bp, which is the optimal size for mutation detection by DHPLC. Larger amplicons also can be used for mutation detection, but the sensitivity of technique decreases with the increasing size. To create homoduplex and heteroduplex molecules, the polymerase chain reaction (PCR) fragments are denatured followed by gradual reannealing such that in the presence of a heterozygous sequence change, "wild-type" and "mutant" sense fragments reanneal with both "wild-type" and "mutant" antisense fragments creating homoduplex and heteroduplex molecules (**Fig. 1**). Homoduplex and heteroduplex molecules are separated on the DHPLC column under partially denaturing conditions (increased temperatures) that cause the heteroduplex molecules to be significantly more denatured than the homoduplex molecules, allowing for their better separation. Homoduplex and heteroduplex molecules bind with differing affinities to the column and elute differently in the presence of an increasing acetonitrile gradient, with heteroduplex molecules eluting earlier. The eluted DNA is detected with an ultraviolet lamp and a chromatogram is generated electronically. A sample with no sequence change will produce only homoduplex molecules, whereas a sample with a sequence change will produce both homoduplex and heteroduplex molecules, with the heteroduplex molecules showing up as an extra peak on the chromatogram. PCR fragments with heteroduplex peaks can be sequenced to determine the exact sequence change present.

The sensitivity of mutation detection by DHPLC is estimated to be between 96 and 100% and very closely matches the sensitivity of direct sequencing ***(1,3,4)***. For this reason, mutation detection by DHPLC is now widely used in both the research and clinical settings ***(5)***.

1.2. Genotyping by DHPLC

The DHPLC instrument can also be used for genotyping of single-nucleotide polymorphisms (SNPs). DNA sequence fragments that differ at a single base pair position can be distinguished on the DHPLC because of the differing hydrophobicities of different base pairs that can cause a change in their elution

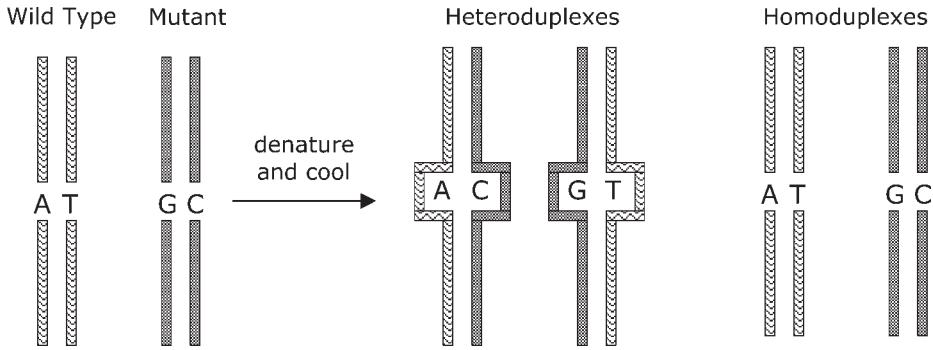


Fig. 1. Heteroduplex and homoduplex formation.

profile ***(6)***. This characteristic is taken advantage of in genotype applications used on the DHPLC and one that has been used successfully is single-base extension (SBE) genotyping ***(7)***. Single-base extension on the DHPLC (SBE–DHPLC) is performed by an initial PCR of an amplicon with a single-base change to be genotyped followed by an extension reaction using an oligonucleotide that acts as a single-base extension primer. The SBE primer is annealed downstream or upstream immediately adjacent to the SNP to be genotyped in the 5' to 3' direction. Thermosequenase extends the 3' end of the extension primer with the appropriate ddNTP. The primer extends one base only because the ddNTP terminates further extension. Extended products are separated on the DHPLC based on the hydrophobicity of the last base, so although the lengths of the extended products are the same for different alleles, the hydrophobicity of the extended products of each allele will be different.

Another variation of SBE genotyping on the DHPLC is primer extension genotyping, in which a combination of dNTPs and ddNTPs are added to the reaction so that, depending on the allele present, either extension beyond the single-base or just single-base extension occurs ***(8)***. Separation of the extended products then becomes a function of the differing lengths of the two extended products. This review will focus on the protocol for SBE genotyping.

The utility of the DHPLC for genotyping is not as widespread as its mutation detection application. However, the utility and effectiveness of SBE–DHPLC for genotyping purposes has been clearly demonstrated ***(7)***. In our experience genotyping by SBE–DHPLC is a very robust technique and often has worked when other methods of genotyping have failed. It is a very useful methodology for medium scale genotyping projects of approx 500 to 1000 samples.

Fig. 2. WAVE™ System, Transgenomic, Inc.

2. Materials

2.1. Instrumentation

1. The WAVE™ system (**Fig. 2**) from Transgenomic Inc., Omaha, NE, is the most widely used system for DHPLC analysis. The methods described in this chapter pertain specifically to the WAVE Nucleic Acid Fragment Analysis System 3500HT, although are applicable to other WAVE model types (*see* **Note 1**). The 3500HT system is a high-throughput system that allows for analysis of hundreds of samples. It consists of six major components (**Fig. 3**): *degasser*, in which the four buffers (A, B, Syringe Wash Solution [Buffer C], and D) originate their flow; *pump*, which controls the percentage of Buffers A, B, and D that flows through the system and prevents contaminants by way of filters from entering the flowpath; *autosampler and chiller*, which contain two 96-well plate holders and the injection needle and valve; *oven*, which contains the inline filter used to filter out particles larger that 0.5 m, a preheat coil, and the separation cartridge/column (DNASep®); *UV detector* (deuterium lamp), which measures the absorbance of DNA samples at 260 nm by light refracting and splitting into two beams; and a *computer*, which plots the absorbance (*y*-axis) against time (*x*-axis) and depicts the DNA as peaks on a chromatogram. An optional *fragment collector* also may be connected to the computer interface. The fragment collector is used to collect and reanalyze separated fragments.

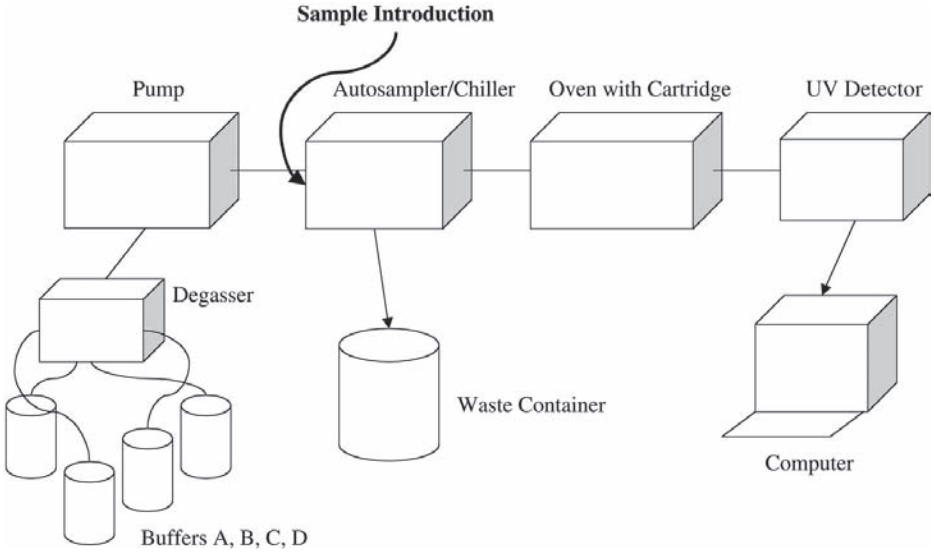


Fig. 3. System flowpath.

The actual flow path of the samples (amplified and denatured PCR fragments) begins with the samples entering the autosampler (**Fig. 3**). Initially, the injection needle, which is connected to the bottom of a glass syringe, is washed with the Syringe Wash Solution (Buffer C) and then drops into the vial/well at which time the syringe plunger goes down and draws a vacuum that removes the sample from the vial/well. The needle moves to the injection port, where the sample is injected into the sample loop. The sample is then carried by buffers to the column in the oven where the DNA fragments are separated based on their structures. As the DNA is eluted from the column by the buffers, it passes through a flow cell where the absorbance of light from the UV detector is measured and plotted on a chromatogram by aid of the computer and associated software.

2. Thermal Cyclers with 96 wells and heated lids are used for PCR reactions and denaturation/reannealing of samples.

2.2. Software

Navigator™ Software Version 1.5.3. (as of this printing) runs the mutation detection application (Transgenomic, Inc., Omaha, NE).

2.3. Column

DNASep® HT Cartridge, cat. no. DNA-99-3710 (Transgenomic, Inc., Omaha, NE) is used for chromatography (*see* **Note 2**).

2.4. DHPLC Buffers/Solvents (see Note 3)

1. WAVE Optimized Buffer A (Transgenomic, Inc., Omaha, NE, cat. no. 553401) or equivalent, 100 m*M* TEAA. Store at room temperature; stable up to 1 mo opened, stable up to 1 yr unopened.
2. WAVE Optimized Buffer B (Transgenomic, Inc., Omaha, NE, cat. no. 553402) or equivalent, 100 m*M* TEAA, and 25% acetonitrile. Store at room temperature; stable up to 1 mo opened, stable up to 1 yr unopened.
3. WAVE Optimized Syringe Wash Solution or Buffer C (Transgenomic, Inc., Omaha, NE, cat. no. 553410) or equivalent, 8% acetonitrile. Store at room temperature; stable up to 1 mo opened, stable up to 1 yr unopened.
4. WAVE Optimized Solution D (Transgenomic, Inc., Omaha, NE, cat. no. 553408) or equivalent, 75% acetonitrile. Store at room temperature; stable up to 1 mo opened, stable up to 1 yr unopened.

2.5. For SBE

2.5.1. PCR Purification for SBE

1. 1X Shrimp Alkaline Phosphatase Buffer (Roche Diagnostics Corp., Indianapolis, IN, cat. no. 1 758 250) or equivalent. Store at –20°C.
2. 1 U of Shrimp Alkaline Phosphatase (Roche Diagnostics Corp., Indianapolis, IN, cat. no. 1 758 250) or equivalent. Store at –20°C.
3. 1 U of *Escherichia coli* Exonuclease I (USB Corp., Cleveland, OH, cat. no. 70073Z) or equivalent. Store at –20°C.
4. Deionized-distilled H_2O (minimum 18 Mohms reading) or HPLC-grade H_2O.

2.5.2. SBE Reaction

1. 1X Thermo Sequenase™ Concentrated Reaction Buffer (Amersham Pharmacia Biotech, Inc., Piscataway, NJ, cat. no. E79000Y) or equivalent. Store at –20°C.
2. Thermo Sequenase Enzyme Dilution Buffer (Amersham Pharmacia Biotech, Inc., Piscataway, NJ, cat. no. E79000Y) or equivalent. Store at –20°C.
3. 2.5 U Thermo Sequenase DNA Polymerase with Pyrophosphatase (Amersham Pharmacia Biotech, Inc., Piscataway, NJ, cat. no. E79000Y) or equivalent. Store at –20°C.
4. 250 µ*M* each ddNTP: ddATP, ddCTP, ddGTP, ddTTP (Amersham Pharmacia Biotech, Inc., Piscataway, NJ, cat. no. 27-2045-01) or equivalent. Store at –20°C.
5. 1 µ*M* extension primer each. Store at –20°C.

3. Methods

3.1. For Mutation Detection

3.1.1. PCR Amplicon Design

The following considerations should be taken into account when choosing amplicons to be analyzed by DHPLC for mutation detection:

1. The optimal size of the amplicons should be 180 to 700 bp.
2. The melting temperature range of the amplicon should be between 52 and 75°C.

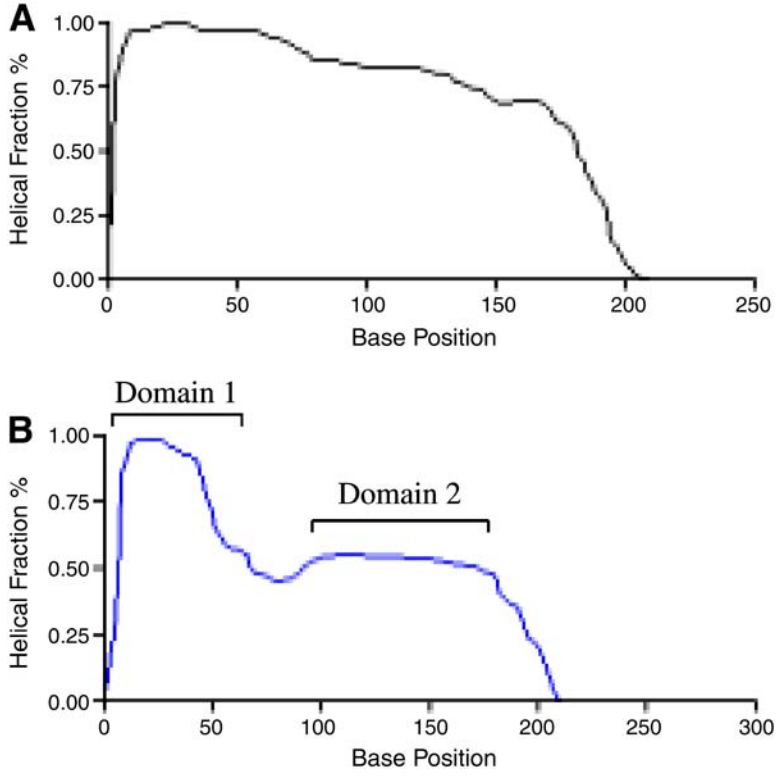


Fig. 4. (**A**) Amplicon with single melting domain. (**B**) Amplicon with two melting domains.

3. Ideally, one should choose an amplicon with one melting domain as opposed to multiple melting domains. **Fig. 4A** shows an amplicon with a single melting domain, and **Fig. 4B** demonstrates an amplicon with multiple melting domains. When the amplicon of interest has multiple melting domains, it may be necessary to break the fragment into smaller amplicons with one melting domain each or incorporate GC clamps (*see* **Note 4**) to PCR primers to even out melting domains within the amplicon thereby obtaining one melting temperature.
4. Amplicon melting profiles are sequence dependent. The GC content within an amplicon also determines the melting profile. The optimal GC content is 48 to 68%.
5. Ideally there should be two degrees or less difference between the melting point T_M of the PCR primers.

3.1.2. Preparation of PCR Samples for Mutation Detection

1. PCRs for subsequent DHPLC analysis are performed using regular touchdown PCR protocols and in 50-µL volumes to allow for sufficient volume injections for DHPLC analysis at various temperatures.

2. Negative control DNA samples should be included for every amplicon being analyzed and positive control DNA samples should be included when available (*see* **Note 5**). A negative control is a sample with no sequence change in the amplicon and a positive control is a sample with a known sequence change in the amplicon being analyzed. A blank (H_2O) control should also be included to check for PCR contamination.
3. After the PCR, samples are denatured and gradually allowed to reanneal to create homoduplex and heteroduplex products. To do this, samples are briefly spun down and denatured and slowly reannealed over 60 min with the following cycling profile: (95°C for 5 min, ramp 95°C → 45°C over 60 min, 45°C for 30 s, hold at 4°C; *see* **Note 6**).

3.1.3. Instrument and Column Preparation

1. The column for sample injections should be installed in the WAVE oven as per manufacturer's instructions and the oven temperature should be set to 50°C for sizing PCR products (*see* **Note 7**).
2. The volumes of buffers A, B, D, and the Syringe Wash Solution (Buffer C) should be checked to make sure that sufficient buffer exists for the number of injections to be performed (*see* **Note 3**). Check waste receptacle, exchange if receptacle is almost full to capacity.
3. Wash the column with 100% Buffer D at flow rate 1.5 mL/min for 10 min. This is performed by entering 100 for %D and 0 for %B and %C and changing the flow rate to 1.5 mL/min on the pump keypad.
4. Wash the syringe five times by pushing the WASH button on the instrument's autosampler.
5. Purge the pumps by setting Buffers A, B, and D to 33% each, flip purge valve in the pump chamber to the "open" position, and press "purge." Purge for 2 min. Press "purge" again and close the purge valve. Purging the pump helps to eliminate air bubbles.
6. Equilibrate the column at 50% Buffer A and 50% Buffer B at flow 1.5 mL/min for 20 min. This step is performed by entering 50 for %B and 0 for %C and %D and changing the flow rate to 1.5 mL/min on the pump keypad.

   Note: **Steps 2** to **6** should be performed once daily before running samples on the instrument. This procedure helps to keep the column and flow path clean and free from impurities. In addition, for optimal instrument and column performance, quality control procedures should be performed at regular intervals (*see* **Note 8**).

3.1.4. Set-Up of Project Defaults

Certain criteria are important to set up as default settings as they pertain to all mutation detection runs. Once these settings are created they need not be entered each time prior to each run.

1. On the Menu Bar, choose *Setup* then *Project Defaults*.
2. In the Equilibrate Cartridge area: check the *Before 1st Injection* box and enter 3 min. This step is necessary for equilibrating the cartridge before the first

injection. Check the *After Temperature Change* box and enter 5 min. Again, it is necessary to equilibrate the cartridge (*see* **Note 9**). Check the *After Gradient Change* box and enter 5 min. This allows for a 5-min equilibration of the column in between changing of the buffer gradient.

3. In the Injection Ordering area: check the *Run in Temperature Order Ascending*. This step allows for samples to be injected in ascending temperature order thereby minimizing the number of times the oven needs to change temperature.
4. In the Clean Options area: check *Normal Clean* (*see* **Note 10**).
5. In the Injection area: Select *Injection Type* ALL. This injection type gives better intensity. In the *Default Injection Volume* enter 7 µL. In the *Feed Volume*, that is, the volume of syringe wash solution injected into the flow path, enter 25 µL when the Injection Type is ALL.
6. *Disable Tray Change Request* is optional. If this is checked, the tray change prompt will not appear when a run is started. This is especially useful when two trays are used for one run.

3.1.5. Creating a New Method

For every amplicon to be analyzed, a method needs to be created. A method contains information or parameters used to run injections. Once a method is created for a particular amplicon, it can be saved and reused. There are three ways to create a method. A method can be created while setting up specific injections on the Injection page or on the DNA page or can be created independently. Guidelines for creating a method independently are detailed in the numbered list below:

1. On the Menu Bar, select File → New Method. Enter a method name. It is helpful to choose a name that includes relevant information such as the name of the gene, exon, type of analysis, etc.
2. Enter the *Application Type* (*see* **Note 11**).
3. Enter the number of *Base Pairs* of the amplicon.
4. Enter the appropriate *Temperature* for analysis (*see* **Note 12**).
5. The default *Injection type* is ALL.
6. The *Clean type* is set at Normal Clean as entered in the Project Defaults.
7. The *flow rate* is the rate at which the buffers move through the system in microliters per minute. The application type will automatically specify the flow rate.
8. The *Percent B* is automatically calculated based on the number of base pairs that is entered.
9. The *Slope* is the amount the %B increases per minute. A slope of 2% increase in Buffer B per minute is the recommended gradient for Mutation Detection ***(9,10)***. The %B should be between the start and stop gradient as indicated on the gradient table.
10. The gradient plot and gradient table are automatically updated when certain parameters including the application type are changed. The gradient plot displays the window of the gradient, that is, it shows the amount of buffer used along the gradient. The horizontal blue line represents the percentage of the buffer(s) indi-

cated in the *Display* field. The blue vertical line indicates where the fragment peak of interest is theoretically predicted to elute under denaturing conditions. The red line (which appears with the Mutation Detection application type only) is a guideline as to where the peak will elute under nondenaturing conditions. The two solid black vertical lines indicate the optimal elution window.

11. The estimated run time is automatically calculated and appears above the gradient table. It should be noted that choosing the application type, *Rapid DNA*, decreases the run time per sample. The *Rapid DNA* application type is the one of choice for the 3500HT system.
12. *Time shift (optional):* The time shift is an adjustment in minutes that moves the elution of the fragment of interest either earlier or later in the gradient. The value of the time shift can be negative (earlier elution) or positive (later elution) with the value between –10 and +10. The time shift actually offsets the gradient by the formula: value x slope. For example, a slope of 2.5%/min and a time shift value of +1.0 min decreases all values for %B (not including clean-off) by 2.5%. The lower percentage of Buffer B results in peaks increasing in retention time. A time shift is recommended if, for example: 1) The peak of interest is eluting too late, which would result in an absence of the peak on the chromatogram. Change time shift default to a negative value such as –1.0, which results in earlier elution. 2) If the peak of interest elutes too early, then change the value to +1.0 min. Essentially, the slope of the gradient changes as a result of a time shift.

3.1.6. Create Sample Sheet (Injection Table)

A sample sheet is a table that specifies the injection order and method type to use for a series of samples to be analyzed. In a sample sheet, information including such variables as sample name, sample location in the tray (vial), method to be used (that links to information such as application type, volume to be injected for each sample, oven temperature, clean type and flow rate) is listed. This needs to be set up for every set of samples to be analyzed and prior to each run.

1. The sample sheet should be set up following the Navigator Software Manual that lists detailed step-by-step instructions.
2. It is recommended that each sample be injected once for sizing (*see* **Note 7**). This is performed using a Sizing application type, that is, DS (double-stranded) Single Fragment (*see* **Note 11**). It is recommended that each sample be injected three times for mutation detection analysis using the three different temperatures calculated for optimal detection of sequence change for the particular amplicon in question (*see* **Note 12** ***[11]***).

3.1.7. Running Samples

1. After all the daily maintenance has been performed (*see* **Subheading 3.1.3.**, **steps 2–6**) and the sample sheet created, run the samples by highlighting specific injections then pressing the run injection button indicated by the green triangle or

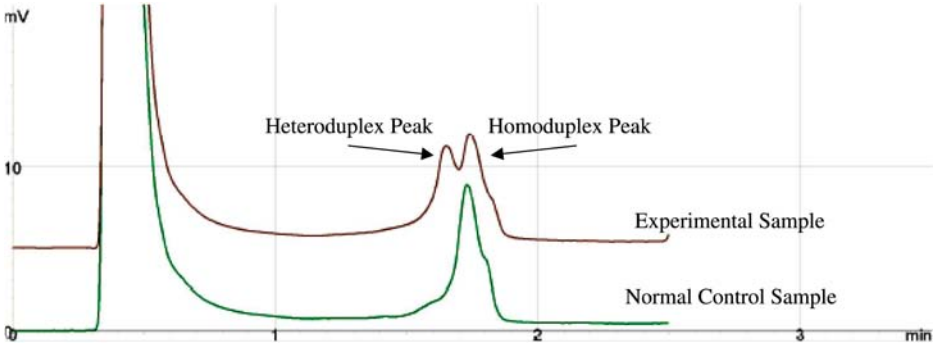


Fig. 5. Normal vs experimental samples indicating homoduplex and heteroduplex peaks.

simply pressing the run injection button when all injections in the sample sheet are to be run.
2. The first 3 min of the run is an equilibration. The equilibration line should be flat at 0 mV. A slight deviation in the line is normal. If the line is not flat at 0 mV the run must be discontinued and the column equilibrated for an additional 10 min.

3.1.8. Analyzing Results

1. As previously mentioned, for all amplicons being analyzed for mutation detection, a normal control (a sample with no sequence change in the amplicon) should be included, and a positive control (a sample with a known sequence change in the amplicon), if available, should be included.
2. Compare the chromatogram of the normal control with the experimental samples for analysis. An absence of a change in the chromatogram between the experimental sample and the normal control indicates no sequence change present in the amplicon of the experimental sample. If a sequence change is present in the amplicon of the experimental sample this will be depicted as an additional peak in the chromatogram as compared to the normal control. The first peak to come off the column represents the heteroduplex product and the second peak that elutes later is the homoduplex product (**Fig. 5**; *see* **Note 13**).
3. Sequence the amplicons of those samples where a change in the chromatogram is observed (*see* **Note 14**).

3.2. For SBE

3.2.1. Preparation of PCR Samples for SBE

1. PCR amplification of the region containing the SNP to be genotyped is performed using regular PCR conditions in a 15-µL volume, with the following exceptions:
   a. Primer concentrations are decreased to 125 n*M* each as excess primer can interfere with the subsequent extension reaction by causing extension to occur from

the PCR primers as opposed to the extension primer. The concentration may be doubled if multiplex reactions are performed or if the amplicon size is larger than usual.

b. dNTP concentrations are decreased to 50 μ*M* each as excess can result in extension beyond the single-base in the subsequent extension reaction. The concentration may be doubled if multiplex reactions are performed or if the amplicon size is larger than usual.

2. The PCR cycling conditions begin with an initial denaturation step at 95°C for 15 min followed by 40 cycles of the following profile: [95°C for 15 s, T_A°C for 15 s (annealing temperature is dependent on primer T_M), 72°C for 30–60 s]. Final Extension at 72°C for 10 min, Hold at 4°C.
3. Check quality and size of PCR products by running 3 μL on a 1.5% agarose gel (*see* **Note 15**). Also include positive and negative controls. Positive controls are samples with known genotypes (*see* **Note 16**). The negative control contains no DNA and therefore should not yield a PCR product. If sizes are correct and yield is adequate, proceed with the purification reaction.
4. The following should be noted with regards to the PCR:
   a. The optimal size of the amplicons should be 150 to 300 bp.
   b. As with all PCRs, when designing primers, avoid 3' end dimers, 3' hairpin loops, and false priming. Primers can be designed using primer analysis software such as Oligo (Version 6.0). Primers can be checked for specificity and to make sure they do not contain polymorphic sites by performing appropriate BLAST (www.ncbi.nlm.nih.gov) and BLAT (www.genome.ucsc.edu) searches.
   c. Multiple SNPs can be genotyped simultaneously by performing multiplex reactions (*see* **Note 17**).

3.2.2. PCR Purification for SBE

1. Purification reactions are performed in a 20-μL volume, and 10 μL of PCR products are used for each reaction. Prepare master mix that consists of the following reagents:

   1 U of Shrimp Alkaline Phosphatase, which removes excess dNTPs from the PCR
   1 U of Exonuclease I to remove excess primers (*see* **Note 18**)
   1X Shrimp Alkaline Phosphatase buffer

   Aliquot 10 μL of master mix to 10 μL of PCR product for each reaction (*see* **Note 19**).
2. Reactions are incubated at 37°C for 45 min followed by inactivation of the enzymes at 95°C for 15 min. Samples can be held at 4°C thereafter.

3.2.3. SBE Reaction

1. SBE reactions are performed in a 10-μL volume. Prepare master mix that consists of the following reagents:

   1X Thermo Sequenase™ Concentrated Reaction Buffer
   250 μ*M* of each ddNTP

1 μ*M* extension primer (*see* **Note 20**)
1.25 U Thermo Sequenase (*see* **Note 21**)

Aliquot 4 μL of master mix to 6 μL of purified PCR product for each reaction. When performing multiplex SBE reactions (*see* **Note 22**), add the additional extension primer(s) to the master mix (also 1 μ*M* concentration) and increase the aliquot of master mix by 0.5 μL for each additional extension primer added. The volume of purified PCR product should be decreased by 0.5 μL (for each additional primer added) as well.

2. The cycling conditions begin with an initial denaturation step at 96°C for 2 min followed by 60 cycles of the following profile: 96°C for 30 s, 55°C for 30 s, 60°C for 30 s. Hold at 4°C.

3.2.4. Denaturing Samples for SBE

1. Denature samples at 96°C for 4 min followed by 4°C; hold before running the samples on DHPLC instrument.
2. In instances in which SBE reactions are pooled prior to running on the DHPLC, a minimum of 8 μL of each individual reaction are combined prior to denaturation (*see* **Note 23**).

3.2.5. Instrument and Column Preparation

1. The column for sample injections should be installed in the WAVE oven as per manufacturer's instructions and oven temperature should be set to 70°C to keep extension products denatured.
2. The volumes of buffers A, B, D, and the Syringe Wash Solution (Buffer C) should be checked to make sure that sufficient buffer exists for the number of injections to be performed (*see* **Note 3**). Check waste receptacle, exchange if receptacle is almost full to capacity.
3. Wash the column with 100% Buffer D at flow rate 1.5 mL/min for 10 min. This is performed by entering 100 for %D and 0 for %B and %C and changing the flow rate to 1.5 mL/min on the pump keypad.
4. Wash the syringe five times by pushing the WASH button on the instrument's autosampler.
5. Purge the pumps by setting Buffers A, B, and D to 33% each, flip purge valve in the pump chamber to the "open" position, press "purge." Purge for 2 min. Press "purge" again and close the purge valve. Purging the pump helps to eliminate air bubbles.
6. Equilibrate the column at 50% Buffer A and 50% Buffer B at flow 1.5 mL/min for 20 min. This step is performed by entering 50 for %B and 0 for %C and %D and changing the flow rate to 1.5 mL/min on the pump keypad.

Note: **Steps 2** to **6** should be performed once daily before running samples on the instrument. This helps to keep the column and flow path clean and free from impurities. In addition, for optimal instrument and column performance, quality control procedures should be performed at regular intervals (*see* **Note 8**).

3.2.6. Set-Up of Project Defaults

Certain criteria are important to set up as default settings as they pertain to all SBE runs. Once these settings are created, they need not be entered each time before each run.

1. On the Menu Bar, choose *Setup* then *Project Defaults*.
2. In the Equilibrate Cartridge area: check the *Before 1st Injection* box and enter 3 min. This is necessary for equilibrating the cartridge prior to the first injection.
3. In the Injection Ordering area: select *Run in Injection Order*.
4. In the Clean Options area: select *Normal Clean*. Do not choose Fast or Active clean (*see* **Note 24**).
5. In the Injection area: select *Injection Type* ALL. This injection type gives better intensity. In the *Default Injection Volume* enter 8 µL for both single and multiplex reactions. For pooled samples, enter 16 µL. In the *Feed Volume*, that is, the volume of syringe wash solution injected into the flow path, enter 25 µL when the Injection Type is ALL.
6. *Disable Tray Change Request* is optional. If this is checked, the tray change prompt will not appear when a run is started. This is especially useful when two trays are used for one run.

3.2.7. Creating a New Method

The *Mutation Detection* application type is used as a template to manually create a new method for SBE–DHPLC runs.

1. On the Menu Bar, select File → New Method. Enter a method name. It is helpful to choose a name that includes relevant information, such as the name of the gene, targeted SNP, type of genotyping assay, and so on.
2. In the opened "Method" window, select/enter the following parameters:
   a. Select *Mutation Detection* as Application Type if it is not shown.
   b. Enter 1.5 mL/min for Flow Rate when using a HT column.
   c. Enter 70°C for Oven Temperature.
   d. Select *Normal* clean as Clean Type as indicated in Project Defaults.
   e. Enter the number of base pairs of the extension primer length or shortest primer length if multiplex SBE is applied. The Navigator Software will calculate the start gradient. This may need to be adjusted if the peak is eluted too early or too late within the run time. The start gradient can be adjusted by performing a time shift (*see* **step 3**). The start gradient also can be adjusted by taking note of what point in the Navigator-calculated gradient the unextended primer elutes at, which is indicative of the percentage B, and the start gradient is adjusted accordingly (usually 1% before the percentage B at which the unextended primer elutes).
   f. Manually change the default settings for the following variables that determine the gradient range and duration:
      Slope (%B/min) 5.0%
      Drop for Loading 5.0%

Loading Duration 0.3 min
Gradient Duration 2.0 or 2.5 min (*see* **Note 25**)
Clean Duration 0.5 min
Equilibration Duration 0.9 min

g. Click Save after entering the parameters but first carefully check all the values of the parameters as sometimes when one parameter is entered or modified, it might also change one or some of the other parameters. If this happens the original values need to be re-entered. This is a result of the software application.

3. *Time shift (optional):* the time shift is an adjustment in minutes that moves the elution of the fragment of interest either earlier or later in the gradient. The value of the time shift can be negative (earlier elution) or positive (later elution). The time shift actually offsets the gradient by the formula: value x slope. For example, a slope of 2.5%/min and a time shift value of +1.0 min decreases all values for %B (not including clean-off) by 2.5%. The lower percentage of Buffer B results in peaks increasing in retention time. A time shift is recommended if, for example, the following occur. 1) The peak of interest is eluting too late, which would result in an absence of the peak on the chromatogram. Change time shift default to a negative value such as –0.5, which results in earlier elution. 2) If the peak of interest elutes too early, then change the value to +0.5 min. Essentially, the slope of the gradient changes as a result of a time shift.

3.2.8. Create Sample Sheet (Injection Table)

A sample sheet is a table that specifies the injection order and method type to use for a series of samples to be analyzed. In a sample sheet, information including such variables as sample name, sample location in the tray (vial), method to be used (that links to information such as application type, volume to be injected for each sample, oven temperature, clean type and flow rate) is listed. This needs to be set up for every set of samples to be analyzed and prior to each run. The sample sheet should be set up following the Navigator Software Manual that lists detailed step-by-step instructions.

3.2.9. Running Samples

The primers should be initially injected individually to check the elution time. This is critical when performing multiplex SBE, as the peaks need to be separated by at least 30 s.

1. After all the daily maintenance has been performed (*see* **Subheading 3.2.5.**, **steps 2–6**) and the sample sheet created, run the samples by highlighting specific injections then pressing the run injection button indicated by the green triangle or simply pressing the run injection button when all injections in the sample sheet are to be run.
2. The first 3 min of the run is an equilibration. Watch to make sure the equilibration line is flat at 0 mV. A slight deviation in the line is normal. If the line is not

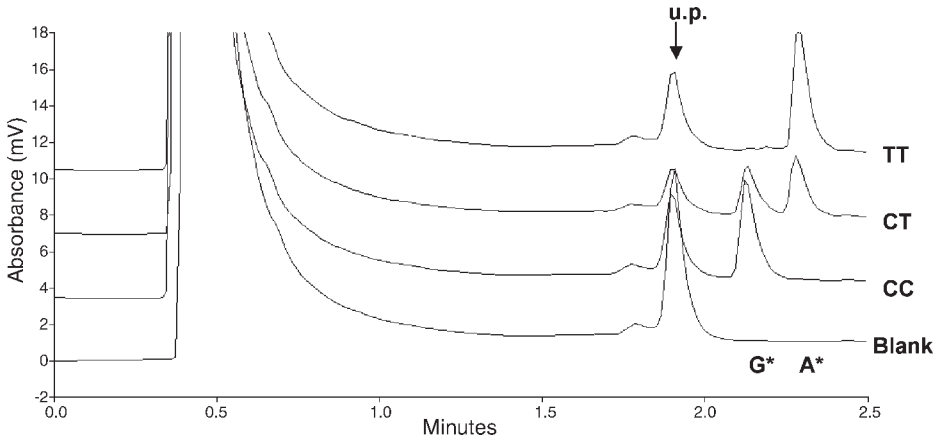


Fig. 6. SBE products. u.p., unextended primer; *reverse extension primer used.

flat at 0 mV the run must be discontinued and the column equilibrated for an additional 10 min.

3.3. Analyzing SBE Results

1. On the *Injection Page*, select the appropriate *Tray Name* for the run that is to be analyzed and click on the *Results* tab. Two charts with *x* and *y*-axes will be displayed. The *x* axis indicates minutes and the *y*-axis represents the absorbance. In the *Injection Table*, under Chart 1 (for graph 1), highlight the blank control as well as the known genotyped control samples. The results of the experimental samples will be compared to these controls and the genotypes determined. Under Chart 2, highlight each sample individually to read the genotype.
2. On the basis of the increasing hydrophobicity of the four bases on the extension products, the elution order is C<G<A<T, that is, the C extension product elutes first and the T extension product elutes last *(7)*. **Figure 6** displays the extension products for a single reaction, whereas **Figs. 7** and **8** show extension products for duplex and triplex reactions, respectively. It should be noted that the elution order of T and A may sometimes be reversed (*see* **Note 26**).

4. Notes

1. Three other WAVE system models also exist. The 3500 is the base model without high throughput capacity. It uses the DNASep cartridge, has a larger mixer than the 3500HT, and does not have an internal accelerator. The smaller volume of the mixer on the 3500HT system allows for an increased flow rate. The 3500A is identical to the 3500 with the addition of the internal accelerator. The third model is the WAVE-MD that is a lower cost and lower throughput system.
2. The DNASep HT cartridge has a larger diameter to accommodate an increased flow rate. The 3500HT system is the only system that can use the DNASep HT

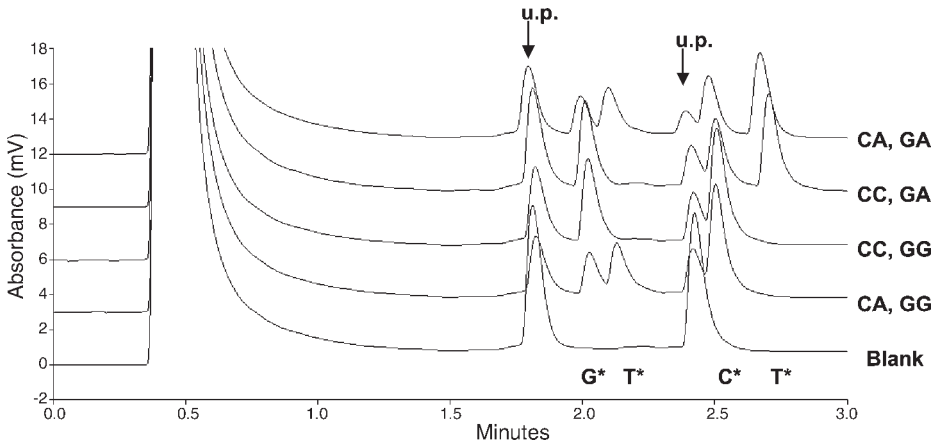


Fig. 7. Duplex SBE products. u.p., unextended primer; *reverse extension primer used.

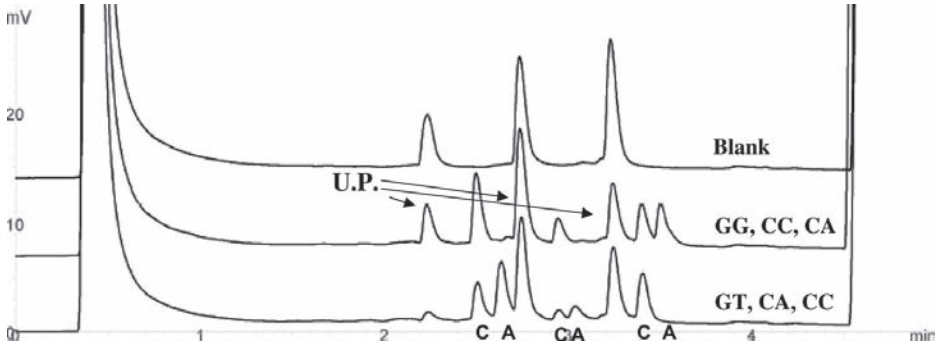


Fig. 8. Triplex SBE reaction. u.p., unextended primer. NOTE: for the first SNP, a reverse extension primer has been used, necessitating reverse complementing of the extended bases for genotype call. For the second two SNPs, forward extension primers have been used.

cartridge. Another analytical column available for DHPLC mutation detection analysis is the DNASep column, which runs at lower throughout than the DNASep™HT cartridge. With proper care and maintenance, the DNASep and DNASep HT cartridges should last a minimum of 10,000 injections.

3. Only ASTM Type 1 (American Society of Testing and Materials) water of 18 MOhm purity with less than 5 ppb Total Organic Content is recommended if buffers are prepared in the laboratory. Do NOT use autoclaved water (for PCRs as well) because there may be metal ions and/or organic contaminants present. Buffers are flammable and irritants. Proper laboratory safety must be exercised

when handling these reagents. Generally, a buffer volume containing approx 1.25 L is sufficient for approx 250 injections. However, this amount is dependent on the duration of each injection, the type of method, etc.

4. Name and Sequence of GC clamps **(2).** The following clamps can be added to PCR primers for amplicon design:

3' End Up	cgggacgc
5' End Up	gcgtcccg
25 bp GCT	cgcccgccgccgcccgccgcccgtc
30 bp GCT	cgcccgccgccgcccgccgcccgtcccgcc
40 bp GCT	cgcccgccgcgcccccgcgcccgtcccgccgccccccgcccg
10 bp GC	cgcccgccgc
15 bp GC	cgcccgccgccgccc
20 bp GC	cgcccgccgccgcccgccgc
3' GC 10 bp	cgggcggggg
3' GC 20 bp	cgggcgggggcggcgggccg
5' GC 10 bp	gcccccgccg
5' GC 20 bp	gcggcccgccgcccccgccg

   If 5' or 3' is not specified, the clamp can be used in forward or reverse direction.

5. A negative control (a DNA fragment with no sequence change) is used to compare the profiles of the experimental samples. A change in peak appearance or number in the experimental samples indicates a potential variant. This must be further analyzed by double-stranded sequencing to determine the exact nature of the sequence variant.
6. PCRs can be performed and stored at 4°C in advance; however, it is essential to denature and reanneal the samples just prior to loading samples on the DHPLC instrument.
7. At 50°C the PCR products are double-stranded and should be sized. All samples, with or without mutations/sequence changes, amplified for a particular region, should have identical elution profiles at 50°C. Sizing is performed to check for the amount of PCR product and its purity before running at the optimal mutation detection temperatures. The PCR product should show up on the chromatogram as a single sharp peak of the expected size. The sizing step on the DHPLC negates the need to run PCR products on agarose gels to check for the presence of amplification products.
8. One of the greatest factors to ensure optimal functioning of the DHPLC instrument is the performance of regular maintenance and quality control procedures. These are detailed below:
   a. The following standards should be run weekly or after approx 500-injection intervals: WAVE DNA Sizing Standard (pUC 18 HAE III digest, 50°C), WAVE Low-Range Mutation Control Standard (56°C), and WAVE High-Range Mutation Control Standard (70°C). Chromatogram peak profiles appear with the reagents. The DNA Sizing Standard is used for size-based DNA fragment separations. Nine fragments with narrow peaks should be detected with

the following sizes in base pairs: 80, 102, 174, 257, 267, 298, 434, 458, 587. Both the 56 and 70°C Mutation Standards form two heteroduplexes and two homoduplexes that assess temperature, elution gradient, and buffer composition. The peak heights should be even and symmetrical and the valley between the two peaks should be deep. Symmetrical peaks with a resolution of approx 50% baseline indicate good resolution and an oven in calibration. Sometimes uneven peak heights or uneven or high valley depths indicate that the oven may need to be re-calibrated or there is loss of resolution. Symmetrical but poorly resolved peaks indicate loss of resolution usually caused by contamination.

If standards deviate from the optimal profiles, it is possible that the column requires a reverse column hot wash with Solution D, as with increasing numbers of injections the column will gradually lose resolution (however, with proper maintenance columns should reach more than 10,000 injections). A reverse column hot wash removes contaminants off of the front side of the column beads. Because the flow direction is reversed, the buildup of contaminants is now washed off the backside of the column. Another possibility of decreased column performance is that the reagents may have degraded and need to be changed.

b. A reverse column hot wash should be performed weekly or after every 1000 injections to clean contaminants off the front end of the column. This is performed by physically reversing the column in the oven and washing the column with 100% Buffer D at 80ºC for 30 min followed by equilibrating at 50% A and 50% B at 50ºC for one hour. Sizing and mutation standards should be run following the hot wash to check the quality of resolution.
c. The inline filter should be changed after every 1000 injections. The inline filter serves to block potential contaminants from reaching the column.
d. An isopropyl alcohol wash should be run to flush the system every 3 mo. This step is done by first washing the column with 100% D for 5 min and then stopping the flow. The column and inline filter are each replaced with a peak union. All solvent lines with the inlet filters attached should be placed in HPLC-grade isopropanol. Purge the pumps for 5 min. Then flush the system for 15 min at a flow rate of 1.5 mL/min with Buffers A, B, and D at 33% each. The autosampler syringe should be washed 20 to 50 times as well. Once the system has been flushed with isopropanol, the solvent lines should be placed in Millipore or equivalent water. The previous purging and flushing steps should be repeated. Equilibrate the system for 20 min with 50% A and 50% B. Note that ALL traces of isopropanol must be removed.
e. It is important to change the solvent inlet filters when they are discolored, that is, no longer white, or slimy every 3 mo. It is possible for the solvent filters to draw a vacuum and cause irregular elution times.
f. If the pressure is unusually high, for example, higher than 1700 (this is checked on the pump display by reading the pressure value) change the inline filter or perform a reverse column hot wash. Often, there may be buildup on either the filter or column, which will cause a pressure increase.

g. The UV lamp energy should be checked weekly. The lamp energy should be equal to or greater than half the initial energy. The wavelength accuracy should be –1 to +1 nm (1). A decreased UV lamp energy will result in lower peak resolution, that is, peaks may be shorter and broader.

9. Equilibrating the cartridge for a minimum of 5 min, between temperature changes, is recommended to help maintain the sharpness and consistency of peaks from the beginning of a set of samples at a given temperature to the end of the set.
10. There are three different clean types. The *Active* clean type runs 100% Buffer D (75% Acetonitrile) through the WAVE System for the amount of clean time specified in the method. The *Fast* clean type injects Buffer D directly into the flow path but is only available for systems with an accelerator. The *Normal* clean type runs 100% Buffer B through the WAVE System for the amount of clean time specified in the method ***(2)***.
11. There are six different application types. The first three types are the most commonly used for mutation detection. *Mutation Detection* is used for creating partially denaturing conditions that will produce heteroduplexes from mixtures of wild-type and mutant fragments. *Rapid DNA*, which can only be run on the 3500HT system, is a faster version of Mutation Detection used for detecting known mutations. *Double-stranded Single Fragment* confirms or determines the size of a single fragment. This type is used for sizing PCR products. *Double-stranded Multiple Fragments* sizes mixtures of fragments over a range of sizes. *Universal Linear* is used for cartridge calibration and general analyses. *Oligo Purification* is a fragment-separation gradient used for purifying DNA within specific size ranges ***(2)***.
12. The appropriate temperature for mutation detection analysis for each amplicon needs to be determined. Each amplicon sequence can be checked by the Navigator software to determine its melting temperature and optimal temperature for mutation detection. The software calculates the optimal temperature for each amplicon based on its sequence and melting temperature. It should be noted that optimal temperatures can range from 48 to 68°C for AT- and GC-rich sequences, respectively ***(9)***. The optimal temperature is the temperature at which partial denaturation of the amplicon occurs that allows for the best separation between homoduplex and heteroduplex molecules. For optimal mutation detection sensitivity, two additional temperatures are used above and below the calculated optimal temperature. Generally, samples are run ±0.5 degrees above/below the calculated temperature as well to capture any changes in the mutation detection peak profile. However, if fragments melt fast as indicated on the Helical Fraction % vs Base Position panel (under the DNA tab), that is, if the helical fraction % curve drops below 50 to 60% within approximately the first 100 bp, then choose smaller temperature increments, that is, ±0.2 degrees.

    In general, the optimal temperature should result in 65 to 95% helicity of the amplicon, that is the portion of DNA that remains double-stranded. Approximately one-third of the amplicon should be above 50% helicity at the chosen temperature.

As per the manufacturer's recommendations, if the GC content is greater than 68%, a mutation detection temperature that is 0.2 degrees higher with each 2% of GC content, than that calculated by the Navigator software, should be used.

13. Some sequence changes may show up as more subtle changes in the chromatogram and therefore the recommendation is to sequence any sample where a deviation from the normal chromatogram pattern is observed. In some instances, the only change in chromatogram between the sample with a sequence change and the normal control may be a decrease in peak height. When a decrease in peak height is obtained, this should be compared to the sizing run (run performed at 50°C when the amplicon is still double-stranded) performed for this sample's amplicon and if the peak height is also observed to be decreased in the sizing run, then this is most likely indicative of a decrease in amplification product as opposed to a sequence change. A sequence change generally shows up as a change in the chromatogram profile in more than one of the three partially denaturing temperatures run for each amplicon, and very often at all three temperatures.
14. Some samples that show more subtle changes in chromatogram profiles, particularly at only one of the three partially denaturing temperatures, may not represent a true sequence change, and may reflect a PCR or DHPLC artifact. It should be noted however that a subtle change, if present at all three temperatures, is very likely to represent a true sequence change. The use of a proofreading Taq polymerase enzyme to generate the amplicons for DHPLC mutation analysis is likely to minimize changes in chromatogram profiles that are due to artifacts as opposed to true sequence changes.
15. When a large number of samples is being genotyped, a selected number of samples can be checked on an agarose gel that is representative for the entire sample set being amplified for genotyping.
16. It is important to simultaneously run controls with known genotypes as this provides a standard peak comparison with the experimental samples. Extended products of the experimental samples can be compared or superimposed to that of the positive control samples. Peak positions will determine the extended base products and therefore the genotype of the sample.
17. Multiplex reactions can be performed in a variety of different ways. This chapter discusses duplex and triplex reactions for genotyping two to three SNPs simultaneously. Although further multiplexing can be performed ***(8)***, it is increasingly more difficult and requires more thorough optimization of conditions. Duplex PCRs are performed when genotyping two SNPs not in close proximity to each other and requires two sets of PCR primers. For triplex PCRs three sets of PCR primers are used. Primer annealing temperatures should be within two degrees of each other for optimal amplification. Also, primer concentrations may need to be adjusted to give equal PCR products as viewed on an agarose gel. If the SNPs to be genotyped are in close proximity to each other they can be included in a single PCR amplicon followed by a multiplex extension reaction. In those instances where multiplex PCRs do not work, PCRs can be performed as individual reactions and then pooled for the subsequent extension reaction.

18. Double the concentration of Exonuclease I if performing duplex PCR, that is, when two sets of PCR primers are used.
19. Also include the blank (H_2O) PCR control in the purification reaction and subsequent SBE-DHPLC reaction. This is added not only to check for the purity of reagents but also to allow for sizing of the extension primer on the DHPLC. Alignment of the blank control reaction with the extension reaction products of the experimental samples on the DHPLC will allow for easy distinction between the extension primer (unextended product) and the single-base extended products (as no extension products will be observed in the blank control reaction; *see* **Figs. 6–8**).
20. The extension primer is designed such that its 3' end lies immediately adjacent to the SNP to be genotyped. The extension primer can be designed to lie either upstream or downstream of the SNP. The sequence surrounding the SNP oftentimes determines which direction to design the extension primer (i.e., upstream/forward or downstream/reverse). The length of the extension primer can range between 18 and 24 bp with an optimal length of 20 bp and aim for a T_M of approx 60°C with approx 50% AT and 50% GC content. When designing the primer avoid areas of repeats and check primers for hairpin loops and primer dimer formation especially at the 3' end of the primer. Single-base changes can be introduced into the primer (usually towards the 5' end) if needed to stabilize the structure, break hairpin loops and so on.
21. Dilute the stock enzyme [32 U/μL] 1:12.8 with Thermo Sequenase™ Enzyme Dilution Buffer. The final concentration should be 2.5 U/μL, and the enzyme solution should be prepared fresh.
22. The following considerations must be taken into account when designing extension primers for multiplex single-base extension. There must be sufficient separation between primers, that is, there must be a minimum of 30 s of elution time separating the primers. This is generally obtained by having primers differ in length by 2 bp or more. GC content as well as the sequence content must be considered. GC-rich primers elute earlier than AT-rich primers. Hydrophobicity of the primers is a factor, for example, elution times are shorter for cytosine and guanine. Elution times may also be adjusted with the addition of GC or AT clamps onto the 5' end of the extension primer.
23. Before pooling reactions, one needs to verify that different extension primers and their extension products can be separated on the DHPLC by sizing aliquots of the extension primers to check their separation. Ideally there should be a separation of a minimum of 30 s between extension primers to allow for clear separation of the corresponding extension products. Pooling can be performed when it is difficult to perform multiplex reactions. Reactions are performed separately and pooled prior to running on the DHPLC. Pooling (as with multiplex reactions) saves on DHPLC run times, DHPLC reagents as well as increases the life span of the DHPLC column.
24. A “normal” clean is the 100% B clean off step containing 25% acetonitrile which prevents column gradient fluctuation for this application, as the gradient condi-

tions for SBE-DHPLC require a low percentage acetonitrile and small increases over time. The "fast" or "active" clean is the 100% D clean off step containing 75% acetonitrile.

25. The gradient duration is modified to 2 min that cuts down the run time. The gradient duration may be increased to 2.5 min or more as needed when multiplex SBE is performed.
26. The elution order of T and A may sometimes be switched at 70°C. Devaney et al. *(7)* reported an elution order of C < G < A < T at 70°C. This elution order has also been observed when a set of four 16-mer heterooligonucleotides differ in a single base at the 3' end *(6)*. We have also observed an elution order of C < G < T < A. Our experimental samples are run at a 1.5 mL/min flow rate using the high-throughput (DNASep HT) cartridge. It is possible that the faster run time as well as the dimensions of the DNASep HT cartridge versus the DNASep cartridge (as used by Devaney et al. *[7]*) have a subtle effect on the elution order of T and A. In addition it is known that retention may be governed not only by the substituted base but also by the immediate sequence context. For this reason the inclusion of positive controls with known genotypes (as determined by an independent method) is very important.

References

1. Oefner, P. J. and Underhill, P. A. (1995) Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC), in ASHG *Annual meeting A2666,* University of Chicago Press, Chicago.
2. *Navigator Software Manual*, Version 1.5.3, © (2003) Transgenomic, Inc., used with permission.
3. Cotton, R. G. (1997) Slowly but surely towards better scanning for mutations. *Trends Genet.* **13,** 43–46.
4. O'Donovan, M. C., Oefner, P. J., Roberts, S. C., et al. (1998) Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics* **52,** 44–49.
5. Frueh, F. W. and Noyer-Weidner, M. (2003) The use of denaturing high performance liquid chromatography (DHPLC) for the analysis of genetic variations: impact for diagnostics and pharmacogenetics. *Clin. Chem. Lab. Med.* **41,** 452–461.
6. Oefner, P. J. (2000) Allelic Discrimination by denaturing high-performance liquid chromatography. *J. Chromatogr. B*. **739,** 345–355.
7. Devaney, J. M., Pettit, E. L., Kaler, S. G., Vallone, P. M., Butler, J. M., and Marino, M. A. (2001) Genotyping of two mutations in the HFE gene using single-base extension and high-performance liquid chromatography. *Anal. Chem.* **73,** 620–624.
8. Wu, G., Hua, L., Zhu, J., Mo, Q., and Xu, X. (2003) Rapid, accurate genotyping of ß-thalassaemia mutations using a novel multiplex primer extension/denaturing high-performance liquid chromatography assay. *British J. Hematol.* **122,** 311–316.
9. Xiao, W. and Oefner, P. J. (2001) Denaturing high-performance liquid chromatography: a review. *Human Mutation*. **17,** 439–474.

10. Taylor, P., Munson, K., and Gjerde, D. (1999) Detection of mutations and polymorphisms on the WAVE™ DNA Fragment Analysis System. Application Note 101. Transgenomic, Inc.
11. Kuklin, A., Munson, K., Gjerde, D., Haefele, R., and Taylor, P. (1997/98) Detection of single-nucleotide polymorphisms with the WAVE™ DNA Fragment Analysis System. *Genetic Testing* **1,** 201–206.

8

Pyrosequencing of Clinically Relevant Polymorphisms

Sharon Marsh, Cristi R. King, Adam A. Garsa, and Howard L. McLeod

Summary

The data generated from the Human Genome Project has led to an explosion of technology for low-, medium-, and high-throughput genotyping methods. Pyrosequencing is a genotyping assay based on sequencing by synthesis. Short runs of sequence around each polymorphism are generated, allowing for internal controls for each sample. Pyrosequencing can also be used to identify tri-allelic, indel, and short-repeat polymorphisms, as well as determining allele percentages for methylation or pooled sample assessment. Assays details for Pyrosequencing of clinically relevant polymorphisms are described in this chapter.

Key Words: Pyrosequencing; genotype; polymorphism.

1. Introduction

The explosion of genetic information available after the completion of the Human Genome Project has proven a goldmine for pharmacogenetic research. Early estimates predicted more than 1.42 million single-nucleotide polymorphisms (SNPs) are present in the human genome ***(1)***. The 121 build of the public SNP repository, dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html) contained 19,888,389 submissions, of which 4,540,241 were validated SNPs. Polymorphisms in coding and control regions of genes can cause significant interindividual variation in the resulting protein function and activity, leading to important differences in disease susceptibility and drug metabolism. This expansion in evaluable SNPs has led to a number of detection methods ***(2,3)***.

Pyrosequencing produces specific sequence data in the form of peaks on a pyrogram. It does not require the presence of a restriction enzyme site, and polymerase chain reaction (PCR) product and internal primer sites can vary in size and position. In addition, it can be used to identify tri-allelic, indel, and short-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

repeat polymorphisms, as well as determining allele percentages for methylation or pooled sample assessment. The availability of sequence directly adjacent to the polymorphisms allows internal quality control checks to be made for each sample. Pyrosequencing is performed on a 96-well platform, and in an average day, more than 3000 individual genotypes can be measured. This method has been used to genotype several clinically relevant polymorphisms ***(4–7)***. Included in these methods are assay details for commonly analyzed clinically relevant polymorphisms in the cytochrome P450 and adenosine triphosphate (ATP)-binding cassette transporter genes CYP3A4 and ABCB1.

2. Materials

2.1. DNA Template

Deoxyribonucleic acid (DNA) from any source can be used in Pyrosequencing assays (*see* **Note 1**). Commonly used kits for DNA extraction, including Gentra and Qiagen, do not inhibit the assay.

2.2. Polymerase Chain Reaction

1. Primer design software (any free or custom bought software is appropriate)
2. 1 to 5 ng DNA template (*see* **Note 2**).
3. PCR mastermix, for example: 30 m*M* Tris-HCl, 100 m*M* potassium chloride, pH 8.05, 400 µ*M* dNTP, and 5 m*M* magnesium chloride (*see* **Note 3**).
4. Hot start Taq polymerase (*see* **Note 4**).
5. DNase- and RNase-free 18.2 mΩ water.
6. DNA oligonucleotides (primers).
7. Unskirted 96-well PCR trays.
8. 96-well sealing film or silicon mat for covering 96-well plates in a thermocycler.
9. Thermocycler with 96-well capacity and heated lid.

2.3. Agarose Gel Electrophoresis

1. Agarose.
2. 50X TAE buffer: for 1 L, add 242 g of Tris base, 57.1 mL of glacial acetic acid, 18.6 g of ethylenediamine tetraacetic acid to 18.2 mΩ water. Store at room temperature. Dilute to 1X with water prior to use.
3. Microwave.
4. Ethidium bromide: 4 µL of 10 mg/mL ethidium bromide/100 mL agarose; add AFTER heating.
5. 6X loading dye: for 100 mL: 30 mL of glycerol, 70 mL of water plus a pinch of bromophenol blue and a pinch of xylene cyanol FF (amount can be varied depending on the desired color). Store at room temperature.
6. Gel apparatus: casting tray, gel tank, lid, and power supply.
7. UV gel documentation system with thermal printer.

2.4. Processing PCR for Pyrosequencing

1. Centrifuge with rotor/buckets to handle 96-well plates.
2. 2X binding buffer: for 1 L, add 1.21 g of Tris, 117 g of NaCl, 0.292 g of ethylenediamine tetraacetic acid to water, pH 7.6 with 1 *M* of HCl. Sterile filter then add 1 mL of Tween-20.
3. Bead mix: 240 µL of streptavidin coated Sepharose beads, 4560 µL of 2X binding buffer and 3600 µL of 18.2 mΩ water per 96-well plate (the magnetic bead processing protocol for a PSQ96 or PSQ96MA is described elsewhere ***[8]***). Excess Sepharose/binding buffer mix can be stored in a glass bottle at 4°C.
4. 96-well plate shaker, e.g., Eppendorf thermomixer (Fisher Scientific, Hampton, NH).
5. Vacuum prep tool and troughs (Biotage, Upsala, Sweden).
6. 70% ethanol in 18.2 mΩ water (*see* **Note 5**).
7. 0.2 *M* NaOH in 18.2 mΩ water.
8. Washing buffer: for 1 L, add 1.21 g of Tris to water, pH 7.6 with 4 *M* acetic acid. Sterile filter.
9. Annealing buffer: for 1 L, add 2.42 g of Tris, 0.43 g of magnesium acetate-tetrahydrate to water, pH 7.6 with 4 *M* acetic acid. Sterile filter.
10. Pyrosequencing primer mix: 12 µL of 0.3 µ*M* Pyrosequencing primer in annealing buffer per well dispensed into a 96-well Pyrosequencing plate (Biotage).
11. Heating block capable of at least 80°C
12. Pyrosequencing plate adaptor set (base and iron) (Biotage).
13. Adhesive sealing film for 96-well plates.

2.5. Pyrosequencing

1. PSQhs96 or PSQhs96A Pyrosequencer with Pyrosequencing 96A version 1.1 or 96MA software or higher. A detailed protocol for the PSQ96 or PSQ96MA has been described previously ***(8)***.
2. PSQ cartridge, capillary dispensing tips or nucleotide dispensing tips, and reagent dispensing tips for hsPSQ96 and hsPSQ96A (Biotage).
3. Pyrosequencing hs reagent kit (Biotage).
4. DNase and RNase free 18.2 mΩ water.
5. Microcentrifuge.

3. Methods

Pyrosequencing is based on sequencing by synthesis. The assay takes advantage of the natural release of pyrophosphate whenever a nucleotide is incorporated onto an open 3' DNA strand. The released pyrophosphate is used in a sulfurylase reaction releasing ATP. The released ATP can be used by luciferase in the conversion of luciferin to oxyluciferin. The reaction results in the emission of light, which is collected by a CCD camera and recorded in the form of peaks, known as pyrograms (**Fig. 1**). When a nucleotide is not incorporated

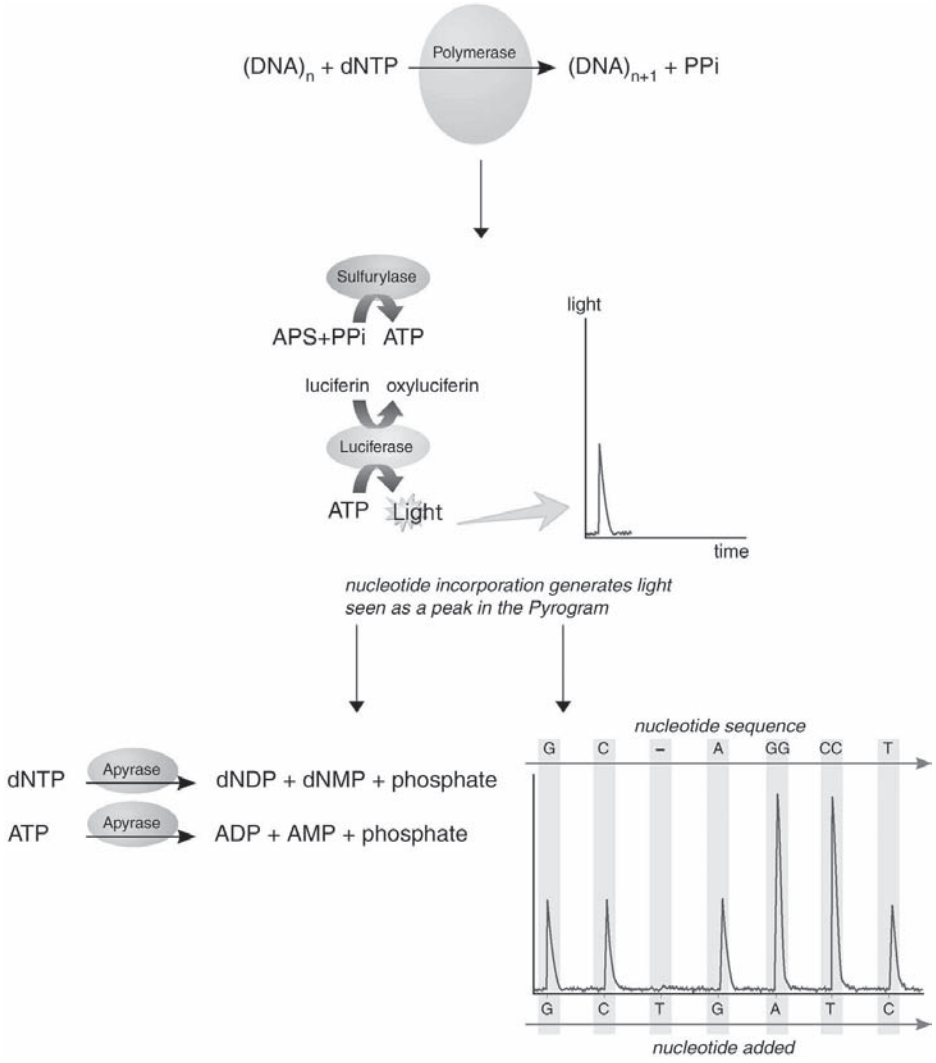


Fig. 1. The Pyrosequencing reaction. A modified ATP is used for the nucleotide dispensations to prevent its direct use by luciferase in the reaction. Modified and published with permission from Biotage AB.

into the reaction, no pyrophosphate is released and the unused nucleotide is removed from the system by degradation through apyrase. This four-enzyme process is performed in a closed system in a single well.

3.1. PCR Primer Design

1. Any primer design software, freely available or custom purchased may be used to design PCR primers for Pyrosequencing. The polymorphism may be in any position of the PCR amplicon from 1 base in from the 3' end of the PCR primer sequence to centered between the primers. SNPs, indels, repeats, etc., do not require specific PCR primer design modifications.
2. Primers should be between 15 and 30 bases long, with an optimum size of 20 bases, ideally with a GC:AT ratio around 50% (although not essential, as you are at the mercy of the location if the polymorphism).
3. Amplicon sizes less than 300 bp are optimal, amplicon sizes 300 to 500 bp give acceptable results, amplicon sizes greater than 500 bp will require optimization during the processing of the PCR product for Pyrosequencing. Amplicon sizes of 100 to 200 bp are suitable for most template sources, including fragmented DNA.
4. Care should be taken to avoid any possible template loops from primers or the single-stranded amplicon doubling back on themselves, as these can lead to background problems during the Pyrosequencing assay (*see* **Note 6**).
5. Optimum primer melting temperature (Tm) is 60°C; however, again, the position of the polymorphism determines the ability to design optimum primers and 50 to 69°C will work. The individual primers should ideally have Tm values within 2°C of each other to allow effective optimization of the PCR.
6. Primer specificity should be checked by screening the primers across available human genome sequence using the NCBI Blast program (http://www.ncbi.nlm.nih.gov/blast/). Extra care should be taken when designing assays for gene family members, e.g., cytochromes, or genes with known pseudogenes, e.g., DHFR, as cross-hybridization of primers can lead to high background, reduced signal and/or false-positive results.
7. One primer needs to be biotinylated at the 5' end. Which primer to biotinylate is dependent on the Pyrosequencing primer orientation.
8. Examples of primers for ABCB1 and CYP3A4 SNPs are shown in **Table 1**.

3.2. Pyrosequencing Primer Design

1. The entire PCR amplicon sequence, including forward and reverse primer sequences, is required to generate the optimum Pyrosequencing primer. This sequence should be pasted into the Pyrosequencing primer design software, freely available through the technical support web site (http://techsupport.pyrosequencing.com/v2/index.asp) to all registered users.
2. Unless multiplexing is required (*see* **Note 7**), the software should be defaulted to find both forward and reverse primers to improve the likelihood of obtaining the optimum primer sequence. The software will list all possible forward and reverse primers by score. A score of 100 is "perfect," scores between 90 and 100 are considered "high." Often "medium" scores yield usable primers, as certain scoring parameters are more critical than others (*see* **Note 8**). Template loops likely to cause background will not affect the overall score but will be highlighted by an asterisk (*see* **Note 9**).

Table 1
Primer Details and Conditions for ABCB1 and CYP3A4 SNPs

Gene	SNP	Orientation	PCR Primers (5'-3')	Anneal	Sequence to analyze
ABCB1	1236C>T	Forward ReverseBiotin Internal	GTGTCTGTGAATTGCCTTGAAGTT **5'-Biotin**/GCATGGGTCATCTCACCATCC TGGTAGATCTTGAAGGG	62	C/TCTGAACCTGAA
	3435C>T	Forward ReverseBiotin Internal	GAGCCCATCCTGTTTGACTG **5'-Biotin**/GCATGTATGTTGGCCTCCTT GGTGGTGTCACAGGAAGA	60	GATC/TGT
	2677G>A/T	ForwardBiotin Reverse Internal	**5'-Biotin**/AGCATAGTAAGCAGTAGGGAGTAACA CTGGACAAGCACTGAAAGATAAGA GATAAGAAAGAACTAGAAGG	62	TG/A/TCTGGGAA
CYP3A4	*1B	Forward ReverseBiotin Internal	AGGACAGCCCATAGAGACAAGG **5'-Biotin**/ATCAATGTTACTGGGGAGTCC CCATAGAGACAAGGGCA	55	A/GGAGA
	*2	ReverseBiotin Forward Internal	**5'-Biotin**/ATCTTCAAATGTACTACAAATCACTGA AACAATCCACAAGACCCCTT TTTGGATCCATTCTTTC	55	TCTC/TAAT
	*3	ReverseBiotin Internal Forward	**5'-Biotin**/GAAGGAGAAGTTCTGAAGGACTCTG CCAGAAACTGCATTGG CGTGGAACCAGATTCAGCAA	65	CAT/CGAGG

3. The orientation of the Pyrosequencing primer will determine the PCR primer to be biotinylated. Forward Pyrosequencing primers require a biotinylated reverse PCR primer, reverse Pyrosequencing primers require a biotinylated forward PCR primer.
4. Examples of internal primers for ABCB1 and CYP3A4 SNPs are shown in **Table 1**.

3.3. PCR Optimization

1. Primer optimization of magnesium concentration and temperature should be carried out in advance for new assays. Ideally a gradient PCR with different magnesium concentrations should be performed, if a thermocycler with a gradient block is available. If a pre-made PCR mix is used, only temperature optimization need be performed (*see* **Note 10**). An example gradient set-up based on a 96-well PCR block with gradient function follows:

   Mastermix (*see* **Note 11**):

   130 µL of Amplitaq Gold PCR mastermix (Applied Biosystems, Foster City, CA)

   Forward primer (10 p*M* final concentration)
   Reverse primer (10 p*M* final concentration)
   13 µL of DNA
   Up to 260 µL with 18.2 mΩ water

   Add 20 µL of mastermix to row of a 96-well plate or 12 0.2 mL tubes and place the gradient block (ensure samples cover a continuous row).

   PCR program (based on an MJ Research (Reno, NV) gradient block):

   93°C for 20 min (or appropriate temperature/time to activate Taq)

   30 cycles of:

   94°C 30 for s
   55 to 72°C 30 for s
   72°C 30 for s

   Then:

   72°C for 5 min
   Store at 4°C.

2. The gradient PCR should be visualized using a 1 or 2% agarose gel. The optimal temperature should give the brightest single band at the appropriate amplicon size. Care should be taken to avoid temperatures where a smeared or multi-band product can be seen as these can increase Pyrosequencing background or reduce specificity if coamplifying a different DNA region. Where several temperatures of equal band intensity are available, the highest temperature should be picked to ensure specificity.

3.4. PCR for Pyrosequencing

1. Care should be taken to avoid contamination. The bench area should be swabbed with 70% ethanol or 5% bleach solution before each PCR set-up and barrier tips should be used for all pipetting steps (if a robot workstation with fixed tips is

used, tips should be cleaned in 5% bleach solution between every DNA dispensation and between every 96-well plate mastermix dispensation).

2. 1 µL of (1–5 ng) DNA (depending on source, *see* **Note 2**) should be dispensed into an unskirted 96-well PCR tray (*see* **Note 12**). At least one well should not contain DNA to act as a negative control (*see* **Note 13**).
3. A 20-µL PCR is ideal for Pyrosequencing; however, if the PCR product is especially strong or wide-peak pyrograms occur, a 10-µL reaction will work well. For a 20-µL reaction based on ABI Amplitaq Gold PCR mastermix (Applied Biosystems, Foster City, CA):

   10 µL of ABI Amplitaq Gold PCR mix
   Forward PCR primer (10 p*M* final concentration)
   Reverse PCR primer (10 p*M* final concentration)
   Up to 19 µL with 18.2 mΩ water
   1 µL of template

4. The PCR plate should be well sealed using a silicon mat or adhesive film. The following PCR program should be run (*see* **Note 14**):

   93°C for 10 min (or relevant temperature/time for Taq activation)

   55 cycles of:

   95°C for 30 s
   X°C for 30 s (based on gradient-derived annealing temp)
   72°C for 30 s

   Then:

   72°C for 5 min
   Store at 4°C.

5. It is possible to directly use the PCR product for Pyrosequencing; however, it is advisable to check the product and the negative control on a 1 to 2% gel to ensure the reaction has been performed successfully and no contamination is present. Contamination is identifiable at the Pyrosequencing stage; however, it is cheaper and faster to run an agarose gel than process and run a contaminated/failed Pyrosequencing plate. 96-well plates should be briefly centrifuged and the lid removed with care to prevent sample aerosol and inadvertent cross-contamination. Typically, 5 µL of the negative control and 5 µL of 5 to 6 wells should give an idea of the success of the PCR. The Pyrosequencing will not be affected by the reduction in volume in these wells. Because of the unusually large number of PCR cycles, some smearing may be visible on a gel, even if the optimum annealing temperature has been used. At this stage the smearing typically does not affect the Pyrosequencing reaction if the PCR primers are specific and the negative control does not contain product.
6. The PCR product can be stored at 4°C until needed. PCR trays should be briefly centrifuged as condensation may occur on the lid, which is a possible source of post-PCR contamination.

3.5. PCR Processing for Pyrosequencing

This protocol assumes the use of a streptavidin/Sepharose bead set-up for Pyrosequencing on a hsPSQ96 or a hsPSQ96A system. The magnetic bead processing method for the PSQ96 or PSQ96MA is described elsewhere *(8)*.

1. A 96-well Pyrosequencing plate containing Pyrosequencing primer mix should be set-up as described in **Subheading 2.4.** (*see* **Note 15**).
2. The small volume readily evaporates, if the set-up time is longer than 10 to 15 min, cover the plate with adhesive film. Primer plates can be aliquoted in advance and stored at 4°C. It is advisable to allow them to reach ambient temperature and briefly centrifuge them before use after storage.
3. Add 70 µL of Sepharose bead mix as described in **Subheading 2.4.** to each well of the PCR product. Replace silicon lid/adhesive film securely.
4. Shake the 96-well plate for 5 min at room temperature. If using the Eppendorf thermomixer, 1400 rpm is the optimum speed. This allows the streptavidin-coated Sepharose beads to anneal to the biotin tag on the PCR primer. Use the plate immediately, if the plate is allowed to sit the beads will settle to the bottom of the wells and will not be accessible to the vacuum tool. If settling has occurred, briefly return the plate to the shaker to disperse the beads.
5. Align reagent troughs, PCR product/bead mix tray and Pyrosequencing primer tray as shown in **Fig. 2** (*see* **Note 16**).
6. With the vacuum switched OFF, shake the vacuum tool tips into clean 18.2 mΩ water. Discard water, refill trough and switch the vacuum on. Place filter tips into trough until all water has been removed (approx 30 s).
7. Place filter tips into the wells containing the PCR/bead mix. Ensure all liquid has been removed from the tray by slightly rocking the vacuum tool can prevent surface tension from causing liquid to remain in the wells. The beads attached to the biotin primer will prevent the PCR product from going through the filters.
8. With the vacuum still on, place the filer tips in the 70% ethanol. Wait a few seconds until a good flow of liquid is seen through the tubing allow the tips to suck up ethanol for 5 s. Repeat with 0.2 *M* NaOH and washing buffer. The NaOH denatures the DNA; therefore, only single-stranded PCR product remains adhered to the filter tips.
9. Switch the vacuum off or remove the vacuum hose from the vacuum tool and place the filter tips into the Pyrosequencing plate containing the Pyrosequencing primer/annealing buffer mix. Residual vacuum will cause the primer mix to be sucked up through the tips so ensure it is fully off. Gently rock the tips in the wells to disperse the PCR product.
10. Place the Pyrosequencing plate onto a heating block at 80°C for 2 min. Ensure the plate sits on the Pyrosequencing plate adaptor with the corresponding lid (or "iron") placed over the plate to prevent evaporation. After 2 min, remove from heating block and place on a bench surface to cool. Once the plate is cool to the touch, cover with an adhesive seal (unless it will be run within 10–15 min) to prevent evaporation. If evaporation has occurred, adding 12 µL of annealing

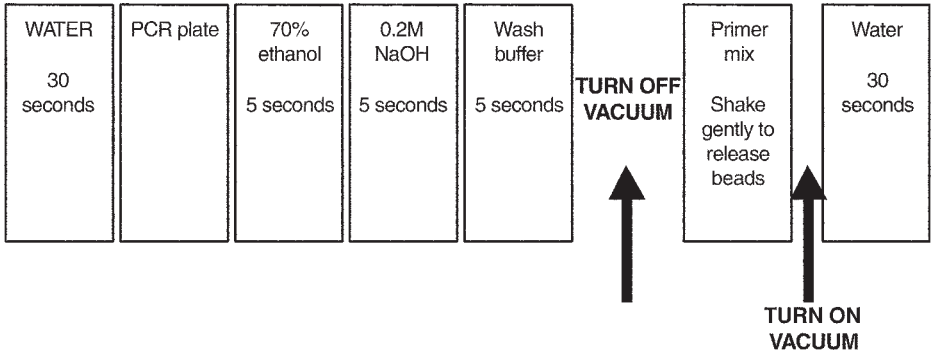


Fig. 2. Reagent layout for processing PCR plates for Pyrosequencing. The same orientation is important for the PCR and Pyrosequencing plates.

buffer will rescue the plate. Covering the plate while it is too hot will cause condensation on the lid, which can lead to cross-contamination of the wells.

11. Processed plates can be stored at 4°C until needed.

3.6. Pyrosequencing

3.6.1. Entering Assay Details

1. Open the Pyrosequencing software. A user name and password is required. This is usually set up with instrument installation. Individual or group-wide passwords can be used.
2. If the assay is not already entered into the software, on the left of the screen click "simplex entry" (*see* **Note 17**). In the menu tree to the right of the simplex entry icon scroll to the top, right click over "simplex entry" and select "new entry."
3. The required fields are a unique name for the assay (usually gene/SNP name) and a sequence to analyze (*see* **Table 1**). Usually, 5 to 6 bases after the SNP position provides enough information for the assay. SNPs should be denoted as, for example, T/C (tri-allelic or tetra-allelic SNPs can also be entered, e.g., G/A/T or G/A/T/C) and indels as e.g., [GATC]. Short repeats should be entered as a series of indels, for instance, [TA][TA][TA]. Clicking "dispensation order" will automatically generate the least amount of nucleotide dispensations required for optimum genotype information. The dispensation order can be edited manually by typing in the dispensation order field, which is useful for troubleshooting problem assays.
4. Select "show histograms" and the predicted pyrogram pattern will be displayed on the right. The default screens show both homozygous patterns and the heterozygous pattern. It is possible to scroll through histograms on the lower panel, useful if multiplex of multiple indels are to be analyzed, etc. Selecting individual or all predicted histograms on the box below the dispensation order and clicking "export" opens the histograms in a browser window where they can be printed or saved.

5. Click "save." At this stage the parameters can no longer be altered, a duplicate set up with a unique name will need to be created for any alterations to the assay.

3.6.2. Entering an SNP Run

1. Select the "SNP run" icon on the far left of the screen.
2. On the menu tree right-click over "SNP run" and select "new SNP run" (*see* **Note 18**).
3. The essential parameters on the setup tab are a unique run name (e.g., gene/SNP/sample set/date) and the active well map. The default plate map is for a full 96-well plate. Individual wells can be selected (hold down control for non-adjacent wells), clicking the "activate wells" button will grey out unused wells. In addition, instrument parameters must be selected from the drop down menu. Usually "instrument parameters" is a default file; however, care should be taken to ensure the appropriate parameters are selected for nucleotide or capillary dispensing tips, as they are not interchangeable. Parameter set up instructions are found with the dispensing tip packaging.
4. The essential parameters on the setup tab are to select the SNP assay by clicking on the drop-down menu under "simplex" and selecting the assay name entered in **Subheading 3.6.1.**, and to fill the plate map by clicking and dragging over the active (white) wells (*see* **Note 19**).
5. Once the run has been set up, click "save." This can be edited post-save, and changes can be re-saved.
6. If multiple plates of the same assay are to be run, on the menu tree right click over the SNP run you have just entered and select "duplicate SNP run." The only parameter necessary is a unique run name.

3.6.3. Individual Plate Run for PSQhs96 and PSQhs96A

1. On the SNP run setup page described in **Subheading 3.6.2.**, click the "view" tab and select "run." This will list the appropriate volumes of nucleotides, enzyme, and reagent needed for the individual run.
2. Set up the cartridge holder as shown in **Fig. 3**. It is essential that all nucleotide/capillary and reagent tips are clean before use. To check for blockages in the nucleotide and reagent tips, fill with 18.2 mΩ water and apply pressure over the top of the tip. Water should squirt from the bottom of the tip. If this does not occur, try filling/emptying the tip several times with water and re-try forcing liquid through. If the tip remains blocked, discard. For nucleotide dispensing tips, do NOT force water through them. The hydrophobic discs may dislodge and prevent the tip from functioning. Rather, ensure the tip has been rinsed several times in water and has been stored in a clean, lint-free environment (*see* **Note 20**).
3. Nucleotides, enzyme, and substrate are sold as a reagent kit. Each vial is clearly labeled. Nucleotides come as a solution, enzyme and reagent are lyophilized and should be resuspended with 18.2 mΩ water before use, the volumes vary per kit and are clearly marked on the labels. The enzyme and substrate both dissolve rapidly and no mixing or shaking is required. Indeed, this should be avoided as

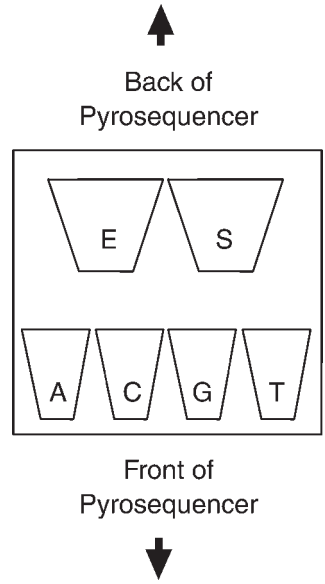


Fig. 3. Reagent and nucleotide cartridge orientation. E, enzyme; S, substrate; A, C, G, and T, nucleotides. A modification of dATP is used to prevent the nucleotide from being a direct source for the oxyluciferase.

air bubbles in the liquid could cause tip blockages or inconsistent dispensation. Unused resuspended enzyme and substrate can be stored at –20°C for future use.

4. If using the nucleotide dispensing tips, the nucleotides should be Microfuged for 10 min and care should be taken to not aliquot from the bottom of the vial in case any precipitate is present which could cause tip blockage. For all dispensing tips it is recommended that non-barrier pipet tips are used as fibers can cause tip blockage.
5. If the capillary dispensing tips are used, the nucleotides should be diluted 1:1 with TE buffer, pH 8.0, and mixed well before use.
6. The nucleotide and reagent dispensing tips should be filled according to the volumes suggested by the software. Nucleotide dispensing tips should be filled by **doubling** the amount suggested by the software. Care should be taken not to pipet air bubbles and to gently angle the liquid down the sides of the tips. Capillary and reagent dispensing tips can allow minute air bubbles without affecting their performance. With nucleotide dispensing tips, it is extremely important to check all of the tips for air bubbles. These can usually be removed by gently tapping the sides of the tips until the air bubbles surface, or, if necessary, dislodging them with a clean pipet tip.
7. A test plate should be run after each cartridge refill. This is extremely important when using the nucleotide dispensing tips, and three or four test plates should be

run in succession to ensure no blockages are present. The substrate reagent-dispensing tip is also prone to blockage if the substrate is allowed to sit in the tip at room temperature for any length of time. To run a test plate: place the cartridge in the pyrosequencer and the test plate in the 96-well plate platform. On the far left of the software screen select the "instrument" tab, then select "instrument" and "manage." Click "test." A warning will appear asking you to check that you have placed the test plate (*see* **Note 21**) into the instrument. Click "ok." The test takes approx 30 s. Remove the plate. In the center there should be six wells with liquid: four nucleotides, a reagent, and a substrate. If there are fewer than six wells with liquid, a blockage has occurred.

8. Remove the adhesive film carefully from the Pyrosequencing plate and place it in the pyrosequencer. Close all levers and click "run" on the plate run set up. The pyrosequencer will now automatically dispense enzyme, substrate and nucleotides in the predetermined dispensation order. The progress of each individual well can be monitored at any time by selecting the relevant well on the 96-well plate map on the screen.
9. To automatically analyze the data once the run has completed, select "analyze all."

3.6.4. Batch Runs Using the PSQhs96A

1. SNP runs should be set up as described in **Subheading 3.6.2.**, saved, and closed.
2. Select the "Batch run" icon on the far left of the Pyrosequencing software, on the menu tree right click over "batch runs" and select "new batch run." One to ten plates can be run in each batch. A unique name for each batch must be provided, and the instrument parameters must be selected for each batch. If barcoded plates are not used, uncheck the "barcode" field.
3. On the far left of the software click on the "SNP runs" icon. From the menu tree, click and drag your SNP runs into the 1 to 10 slots on the batch window.
4. On the top menu bar select "batch" and "setup information." This will open a browser window (may take a few seconds) with the total amount of nucleotides (which should be **doubled** for the capillary dispensing tips), enzyme and reagents needed for the entire batch.
5. The cartridge should be set up as described in **Subheading 3.6.3.** The dispensing tips should be cleaned between every batch and a test plate should be run before every batch.
6. Remove the adhesive film from the Pyrosequencing plates and stack them (check that the plates can be lifted free without sticking to the lower plates, occasional warping may occur, causing plates to stick together, which jams the robotic arm). Place plates in the robot stacker unit. The correct plate orientation is shown on the top of the stacker unit. Ensure the plates lie flat on the base of the stacker unit and are between the grooves. Plate 1 on the Batch set up should be on the top, plate 10 (or the last plate in the batch set up) should be on the bottom.
7. Ensure the stacker unit is firmly pushed into place. The nucleotides will not dispense if the unit is only partially home.

8. Click the "play" icon. Plates will automatically load and be discarded throughout the batch.
9. Plates will automatically be analyzed by the software when run in batch mode. They can be accessed from the batch set up window or from the individual SNP run files.

3.6.5. Analysis of Pyrosequencing Results

1. Once the Pyrosequencing run has been analyzed by the software, the 96-well plate map will be color-coded according to the result. Blue indicates a well in which the pyrogram matches one of the predicted histograms and a genotype can be accurately called. Orange indicates a possible match with a predicted histogram; however, human intervention is required to validate the call. Red indicates a failed well, where no match with a predicted histogram can be found. **Figure 4** shows pyrograms and associated predicted histograms for the triallelic ABCB1 2677 G>A/T polymorphism.
2. The well(s) where no DNA was added in the PCR should automatically be scored failed (*see* **Note 22**). Nonspecific peaks in the negative control(s) may be evident. These are likely to be caused by looping of the internal primer and can aid troubleshooting assays by identifying whether the internal primer is the culprit for background peaks.
3. Samples checked (orange) for human intervention can be edited by clicking on the specific well and opening up the predicted histograms from the "histogram" tab on the right. If a genotype consensus is reached the sample call can be manually edited by right-clicking over the genotype above the pyrogram. Genotypes can be selected and pass/check/fail can be altered. The well on the plate map will show a dark circle, indicating that manual editing has taken place.
4. The data can be exported as a report, as a tab delimited file, or an XML file. Custom export options are also available. The export function can be accessed by selecting "report" and then saved as the appropriate file type. Selected wells or the entire plate can be saved/exported. Pyrograms (all or selected) can also be saved or printed, up to 6 per page (*see* **Note 23**).

4. Notes

1. Pyrosequencing has been successfully performed on DNA from cell lines, blood, serum, plasma, paraffin-embedded tissue frozen tissue, and whole genome-amplified product. In addition, complementary DNA from various sources has also been successfully pyrosequenced.
2. The actual starting concentration of DNA depends on the quality of the template. DNA extracted from blood is highly accessible for PCR and consequently 0.5 to 1 ng can produce reliable, reproducible product. DNA from plasma, serum, frozen tissue, and whole genome-amplified methods tend to be fragmented and more template may be necessary for optimum PCR. A test in advance of serial dilutions of the template DNA should be performed with the PCR primers to find the appropriate concentration that gives a clean high-yield PCR product.

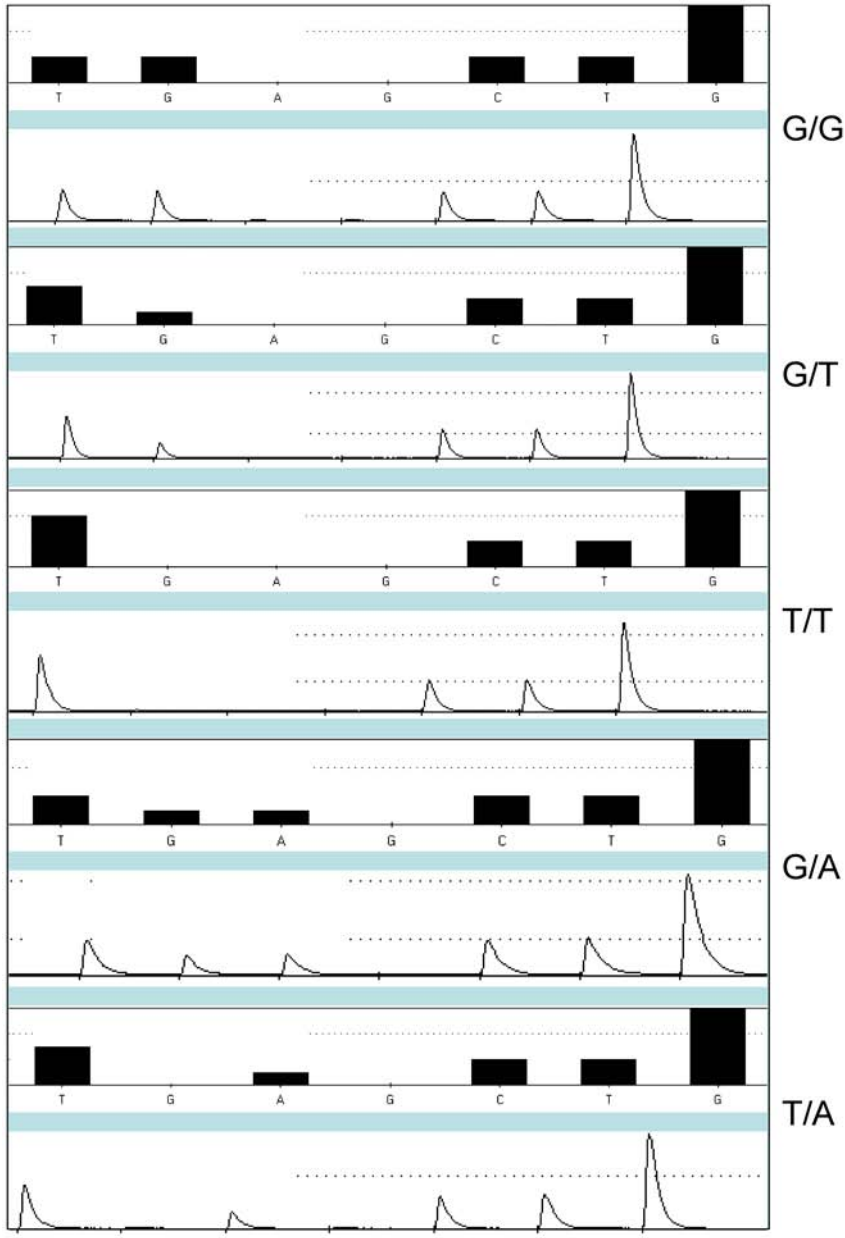


Fig. 4. Predicted histograms and pyrograms for ABCB1 2677 G>A/T genotypes.

3. Premade mixes of buffer, magnesium, dNTPs and Taq polymerase are recommended as they provide consistent results and minimize pipetting errors.
4. Non-hot start Taq is also suitable; however, primer dimers are less of a problem with hot start Taq and this is recommended.
5. All solutions should be made using 18.2 mΩ water. Solutions other than the NaOH and 70% ethanol should be sterile filtered prior to the addition of Tween-20. 10X washing buffer, annealing buffer, and NaOH can be made and stored at room temperature for dilution to the working concentrations. All solutions can be stored at room temperature.
6. Problem template loops will also be flagged in the Pyrosequencing primer design software.
7. This protocol is based on simplex assays; however, multiplexing with up to 3 internal primers can be performed, either from the same PCR product or different PCR products. The primer design software can only determine one internal primer at a time, often the first choice primers for each will not be useful in a multiplex assay where the combined sequence to analyze is best designed to generate unique SNP dispensations. In addition, the orientation of the primers is vital for multiplex assays as only one PCR primer can be biotinylated.
8. Critical scoring parameters on the Pyrosequencing primer report*:*

   Mispriming: If the internal primer can anneal to multiple positions within the amplicon the 3' ends of the annealed region can incorporate nucleotides leading to incorrect genotype calls or unacceptable background.

   Duplex Formation: If the internal primer can dimerize with itself, as for the mis-priming, unacceptable background may result, or reduced signal intensity owing to sub-optimum primer annealing.

   Hairpin Loop: If the primer forms secondary structures the amount of primer available for the reaction is diminished and reduced signal can result.

   Template Loop: Loops of more than approx 4 to 5 GC-rich regions will be flagged by an asterisk and should be avoided. Loops less than 4 bases should also be avoided if possible to reduce the likelihood of background.

   Noncritical scoring parameters on the Pyrosequencing primer report:

   Repeated Base at SNP Sequence: This reduces the score if there are a string of identical bases around the SNP. This is not something that can be controlled or optimize for as the SNP position is not moveable. Typically the pyrograms can accommodate as many as three bases in a row with no problems. Four-six bases may be difficult to read manually as the scale will be affected. More than six repeated bases is not recommended because distinguishing the peak heights become very difficult.

   Primer Length: Primer lengths longer than 15 bases give a reduced score based on the expense of longer primers. The length of the primer is not critical to the reaction.
9. If an appropriate Pyrosequencing primer cannot be found as the critical scoring parameters are flagged, it is possible to “trick” the software to improve the search. As the software will only look five bases to either side of the SNP for a suitable

primer, entering a fake SNP five bases before or after will extend the region searched. This may help to overcome mispriming and dimer problems. To eliminate template loops, adjusting the 5' end of the PCR primer that would cause the loop will help, e.g., shifting the primer two to three bases to the left or right, or trying a PCR primer in a slightly different region. As only one primer is likely to cause the loop problem, if a primer in the opposite orientation is available (even if not the highest score), this is often the easiest solution.

10. Premade PCR mixes are usually a fixed magnesium chloride concentration. If primer conditions are not optimized through temperature alone, extra magnesium chloride may be added to the PCR mix. In addition, problem assays may be improved by the addition of 5 to 10% dimethyl sulfoxide or 1 *M* Betaine. This will not affect the Pyrosequencing.
11. The mix is for 13 samples, allowing one extra sample for pipetting discrepancies.
12. If a larger volume of DNA is necessary, adjustments can be made to the PCR mastermix (reducing the water volume), or DNA may be dispensed into the plate and allowed to dry down overnight at room temperature. The DNA is reconstituted once the PCR mastermix is added.
13. For multiple primer sets/plate, at least one negative control/primer set should be included.
14. Fifty-five cycles are run to ensure all primers and nucleotides are exhausted and not available to cause background during the Pyrosequencing. If wide peaks occur in the program, reducing the number of cycles to 40 may help to prevent these.
15. Multiple assays can be run/96-well plate, indeed, each well could contain a different internal primer. The wells corresponding to the negative controls from the PCR set up should contain internal primer, as this is a valuable trouble-shooting method for program background issues.
16. A workstation platform is available from Pyrosequencing, which holds the reagent troughs and plates in specified positions. Any method to hold the reagent troughs stationary is appropriate, for example, rigid plastic tip box lids.
17. If a multiplex assay is to be set up, select the "multiplex entry" icon, right click over "multiplex entry" on the menu tree and select "new entry." Type in the three separate dispensation orders for each internal primer. The computer generated dispensation order will give a combined dispensation for the three SNPs. The field requirements here are the same as for the simplex entry except two or three sequences to analyze may be entered.
18. The menu tree for SNP runs can be organized into folders so multiple users can easily access their files. If this has been done, right click over the relevant folder and select "new run."
19. Each well can contain a different simplex/multiplex entry if desired, simply select the entry and click in the appropriate well until all active wells are filled.
20. Pyrosequencing provides specific storage boxes for the tips with the instrument, and more are available from the company if required.
21. To save on plate costs, attach adhesive film to the top of the test plate. The dispensation will occur on the film, rather than in the wells and this can be wiped off and the plate can be re-used.

22. If multiple primer sets are used/plate, the negative controls for each primer set should be checked for contamination.
23. The report structure is available in forms readily transferable to most database/spreadsheet systems.

Acknowledgments

The assistance of Derek Van Booven is greatly appreciated. This work is supported by the NIH Pharmacogenetics Research Network (U01 GM63340); http://pharmacogenetics.wustl.edu.

References

1. Sachidanandam, R., Weissman, D., Schmidt, S. C., et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933.
2. Kwok, P. Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* **2,** 235–258.
3. Freimuth, R. R., Ameyaw, M.M. , Pritchard, S. C. , Kwok, P. Y., and McLeod, H. L. (2004) High-throughput genotyping methods for pharmacogenomic studies. *Curr. Pharmacogenom.* **2,** 21–33.
4. Ahluwalia, R., Freimuth, R., McLeod, H. L., and Marsh, S. (2003) Use of pyrosequencing to detect clinically relevant polymorphisms in dihydropyrimidine dehydrogenase. *Clin. Chem.* **49,** 1661–1664.
5. Mathijssen, R. H., Marsh, S., Karlsson, M. O., et al. (2003) Irinotecan pathway genotype analysis to predict pharmacokinetics. *Clin. Cancer Res.* **9,** 3246–3253.
6. Saeki, M., Saito, Y., Jinno, H., et al. (2003) Comprehensive UGT1A1 genotyping in a Japanese population by pyrosequencing. *Clin. Chem.* **49,** 1182–1185.
7. Garsa, A., S. Marsh, S., and McLeod, H. L. CYP3A4 and CYP3A5 genotyping by pyrosequencing. *BMC Medical Genetics.* In press.
8. Rose, C. M., Marsh, S., Ameyaw, M. M., and McLeod, H. L. (2003) Pharmacogenetic analysis of clinically relevant genetic polymorphisms. *Methods Mol. Med.* **85,** 225–237.

9

Kinetic Fluorescence-Quenching Detection Assay for Allele Frequency Estimation

Ming Xiao and Pui-Yan Kwok

Summary

The analysis of human genetic variations, such as single-nucleotide polymorphisms (SNPs), has great applications in genome-wide association studies of complex genetic traits. We have developed an SNP genotyping method based on the primer extension assay with fluorescence quenching detection. The template-directed dye-terminator incorporation with fluorescence quenching detection (FQ-TDI) assay is based on the observation that the intensity of fluorescent dye R110- and R6G-labeled acycloterminators is universally quenched once they are incorporated onto a deoxyribonucleic acid (DNA) oligonucleotide primer. By comparing the rate of fluorescence quenching of the two allelic dyes in real time, we have extended this method for allele frequency estimation of SNPs in pooled DNA samples. The kinetic FQ-TDI assay is highly accurate and reproducible both in genotyping and in allele frequency estimation. Allele frequencies estimated by the kinetic FQ-TDI assay correlated well with known allele frequencies, with an r^2 value of 0.993. Applying this strategy to large-scale studies will greatly reduce the time and cost for genotyping hundreds and thousands of SNP markers between affected and control populations.

Key Words: Single-nucleotide polymorphism; allele frequency; fluorescence quenching.

1. Introduction

A number of areas exist in pharmacogenomics research where it is desirable to determine the relative abundance of the two alleles of a gene or marker. For example, humans have a number of mast cell tryptase genes, and the number of copies of beta-tryptase is associated with asthma ***(1)***. Routine genotyping can only determine the presence or absence of the beta-tryptase gene but cannot determine the number of copies of beta-tryptase gene found in the genome. Another scenario concerns the duplication or deletion of genes in tumor cells caused by somatic mutation. Being able to determine whether a gene is dupli-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

cated or deleted is very helpful in studying cancer progression. An emerging area of research has to do with regulation of gene expression. Here, the relative abundance of alleles of a genetic marker in coding sequences can be used to monitor the expression differences between the allelic genes from the two chromosomes in the cell. In extreme cases, one can estimate the allele frequencies of genetic markers in pooled samples in genetic association studies ***(2)***.

Although a number of quantitative assays are available for allele frequency estimation, most are difficult and expensive to develop or implement. In this chapter, we describe the most versatile and least expensive assays for allele frequency estimation, namely, the template-directed dye-terminator incorporation assay with fluorescence quenching detection (the FQ-TDI assay ***[3]***). The FQ-TDI assay is a real-time homogeneous primer extension assay based on two principles, namely, that deoxyribonucleic acid (DNA) polymerase catalyzes the allele-specific incorporation of a dye-terminator at the polymorphic site and that the fluorescence intensities of a fluorescent dye decreases significantly when it is incorporated into primers.

The TDI assay procedure is quite simple. First, the DNA template containing a SNP site is amplified by polymerase chain reaction (PCR). After the excess PCR primers and deoxynucleotides (dNTPs) are degraded by exonuclease I and shrimp alkaline phosphatase, a SNP primer designed to hybridized one base upstream of SNP is allowed to anneal to the target DNA in the presence of DNA polymerase and dye-labeled acycloterminators. The SNP primer is extended one base by the acycloterminator complementary to the allele present on the target DNA. By determining which terminator is incorporated, one can infer the genotype of the target DNA sample. The primer extension has been used in various format for SNP genotyping and proven to be highly specific and sensitive ***(4,5)***.

If the TDI assay is monitored in real time, one can calculate the rate of incorporation, which is proportional to the amount of allelic target present. Therefore, one can estimate the allele frequency in the sample by determining the relative incorporation rate of the two terminators. A simple way to monitor the progress of the TDI assay is to take advantage of the physical phenomenon of fluorescence quenching. The fluorescence of some conjugated fluorescent dyes is sensitive to their local environment, such as the presence of DNA nearby. The fluorescence quenching caused by DNA–dye interactions has been reported for many different fluorescent dyes ***(6,7)***. We found that R110- and R6G-labeled acycloterminators were heavily quenched upon incorporation into oligonucleotides, especially those containing Gs within 10 bases of the incorporation site. **Figure 1** shows the real-time fluorescence intensity profiles of four representative samples tested for SNP marker rs154162 during the primer extension step of the TDI assay. The fluorescence readings correspond to the

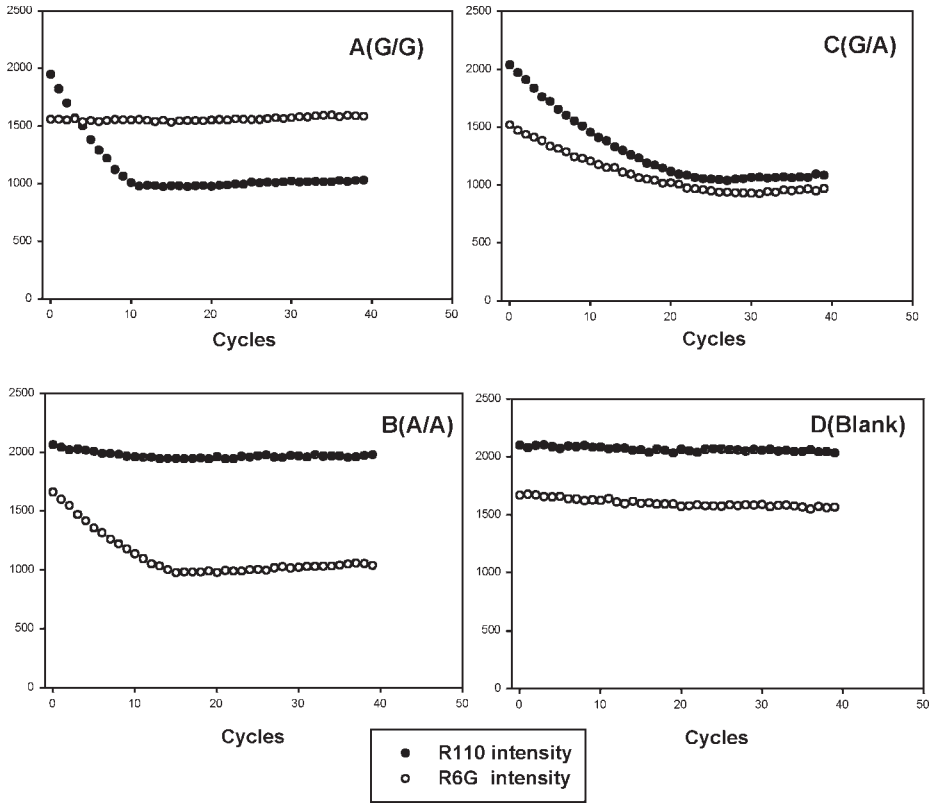


Fig. 1. The real-time fluorescence intensity profiles of four representative samples tested for SNP marker rs154162 during thermal cycling of the primer extension step of the TDI assay. R110 fluorescence is represented by solid circle and R6G fluorescence is represented by open circle. **(A)** shows the intensity profiles of a G/G homozygous sample; **(B)** an A/A homozygous sample; **(C)**, a G/A heterozygous sample; and **(D)**, a negative control.

emission maxima for R110-G (525 nm) and R6G-A (535 nm) acycloterminators are normalized by multicomponent analysis to account for the crosstalk between the two dyes. In **Fig. 1A**, the G/G homozygous sample directs the incorporation of R110-G (but not R6G-A) and the fluorescence reading shows a progressive drop in R110 fluorescence (filled circles) but no change in R6G fluorescence (open circles) as thermal cycling proceeds. The homozygous A/A sample in **Fig. 1B** shows that R6G-A is incorporated (but not R110-G) with a drop in R6G fluorescence but no change in R110 fluorescence. The heterozygous G/A sample in **Fig. 1C** shows that both R110-G and R6G-A are incorpo-

rated, with both R110 and R6G fluorescence dropping as the reaction proceeds, albeit at a slower rated than those observed in homozygous samples. In contrast, the fluorescence intensity profile of a negative control sample (**Fig. 1D**) shows no change for both dyes because neither is incorporated.

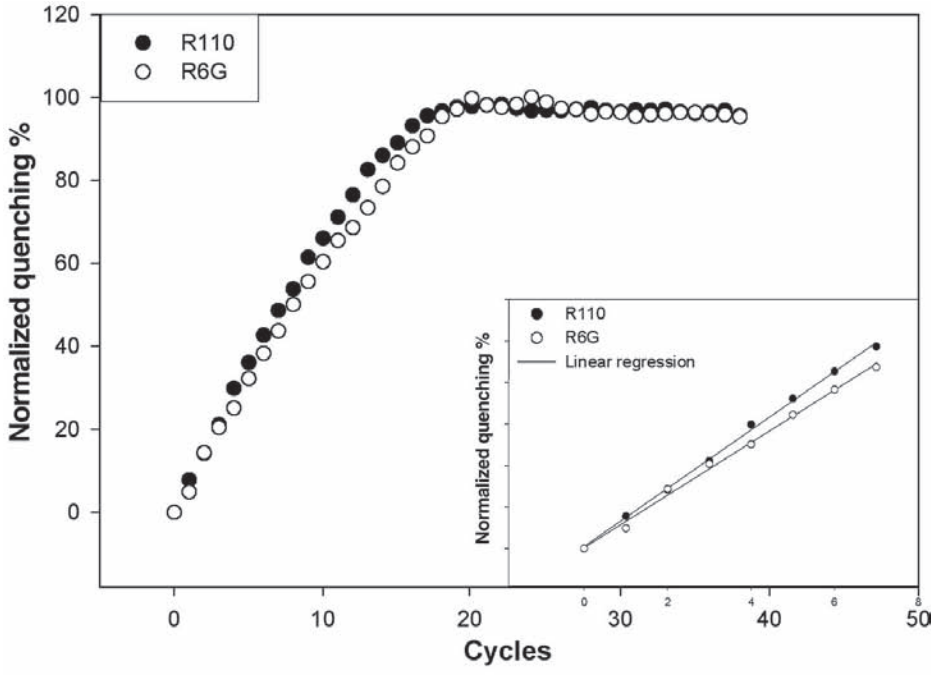
The rate of change of the R110 and R6G fluorescence intensities is proportional to the amounts of the allelic target present in the reaction, with the initial slope of change for the heterozygote being approximately half of that for the homozygote.

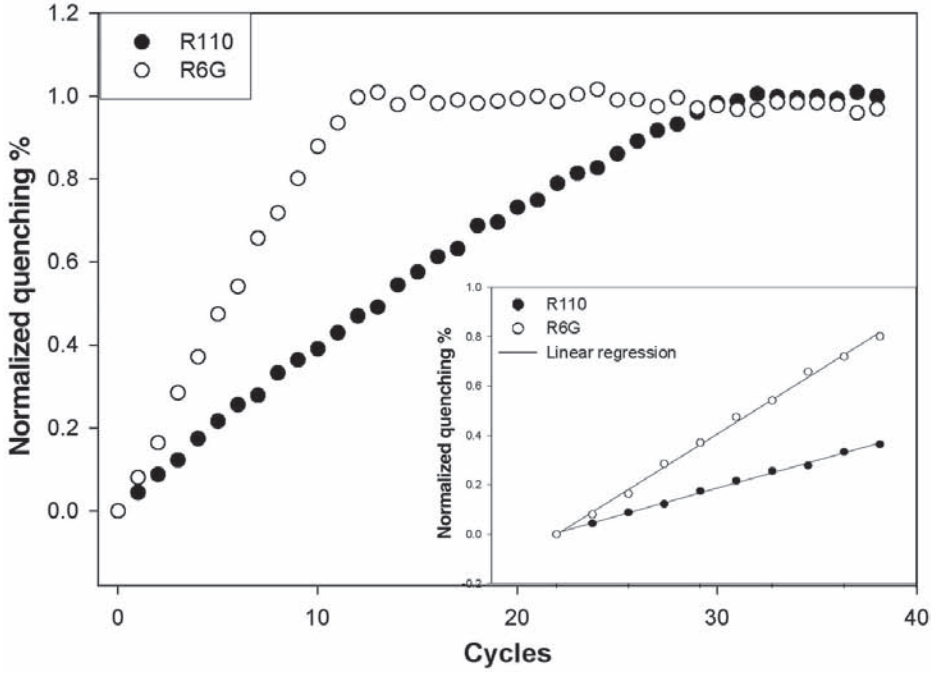
Because the amounts of dye terminators used in the primer extension reaction are fixed by the manufacturer, the larger the number of target DNA molecules containing a particular allele that is present, the faster the dye terminators are used. The initial rate of change in fluorescence intensity (the steepness of the slope) of a dye-terminator is therefore a reflection of the amount of DNA molecule containing the allele corresponding to the dye-terminator incorporated. By monitoring the fluorescence profile during the primer extension reaction, one can calculate the initial rate of change of fluorescence intensity. Comparing the initial rate of change of fluorescence intensity between a test sample and a heterozygous reference sample allows one to infer the relative amounts of the alleles found in the test sample. **Figure 2** illustrates a typical reaction performed to determine allele frequency of pooled DNA samples. The top panel is a normalized real-time quenching curve for a heterozygous sample of rs922365. The linear regressions of the first eight cycles are shown in the inset, and the ratio of the two slopes is calculated as 0.93 (allele 2 [R6G]/ allele 1 [R110]), which indicates that the AcycloPol enzyme incorporates R110-G slightly faster than R6G-A. The ratio of two slopes (S_2/S_1) of the pooled sample in the bottom panel of **Fig. 2** reflects the relative amount of allele present in the pooled sample. The higher the value of the slope, the more copies of this particular allele are present in the pooled sample. To calculate the absolute allele frequencies, one needs to correct differential incorporation efficiency of two dye-acyloterminators by the polymerase using a heterozygous sample $C = (V_1/V_2)$ as indicated in Eq. 1 ***(8)***.

$$p_1 = 1/[1 + C(S_2/S_1)] \quad (1)$$

where p_1 is the frequency of allele 1 and V_1/V_2 is the ratio of the slopes for alleles 1 and 2 of the heterozygous reference sample.

Fig. 2. *(opposite page)* The basis of determining allele frequency for pooled DNA samples. The top panel is a normalized real-time quenching curve for a heterozygous sample of rs922365. The linear regression of the first eight cycles is shown in the inset. The bottom panel is the quenching curve for a mixture containing 75% A allele (R6G).


R110
R6G
Normalized quenching %
Cycles
R110
R6G
Linear regression
Normalized quenching %


R110
R6G
Normalized quenching %
Cycles
R110
R6G
Linear regression
Normalized quenching %

2. Materials

2.1. Polymerase Chain Reaction

1. PlatinumTaq DNA polymerase with 10X PCR buffer and 50 m*M* $MgCl_2$ solution (Invitrogen, Carlsbad, CA).
2. dNTP mixture: 2.5 m*M* dATP, 2.5 m*M* dCTP, 2.5 m*M* dGTP, and 2.5 m*M* dTTP.
3. PCR primers were designed using modified Primer 3 ***(9)***. Working solutions of the PCR primers (P1 and P2) are prepared at concentration of 2.5 μ*M* (*see* **Note 1**).

2.2. Degradation of Excess PCR Primers and dNTPs

1. Exo-Sap (Exonuclease I and shrimp alkaline phosphatase; USB, Cleveland, OH).

2.3. Primer Extension

1. SNP detection kit (AcycloPrime), including 10X reaction buffer, AcycloPol enzyme (PerkinElmer, Boston, MA).
2. R110-acycloterminators (aATP, aCTP, aGTP, aTTP) and R6G-acycloterminators (aATP, aCTP, aGTP, aTTP; PerkinElmer). Unlabeled acycloterminators (aATP, aCTP, aGTP, aTTP; New England Biolabs, Boston, MA).
3. SNP-specific primers were designed with a custom made program ***(9)***. The SNP primer stock solution is prepared at 10 μ*M* concentration (*see* **Note 2**).

2.4. Instrument

1. 96-well or 384 well optic plates (Applied Biosystems, Foster City, CA).
2. Thermocycler.
3. Applied Biosystems 7700 or 7900 Sequence Detector.

3. Methods

3.1. Pool DNA Sample Preparation

1. DNA concentration is quantified by using both the absorbance at 260 nm and a DNA specific fluorescence dye, PicoGreen (Molecular Probes, Eugene, OR) according to the manufacturer's instructions.
2. Construct the population pool samples by adding equal amounts of DNA from each of the 32 individuals to yield a pooled sample containing a final DNA concentration of 10 ng/μL.

3.2. PCR Amplification

1. Amplify genomic DNA (10 ng) in 5 μL reaction mixtures containing 0.5 μL of PCR primers, 0.25 μL of $MgCl_2$, 0.5 μL of 10X PCR buffer, 0.2 U of Platinum Taq polymerase, and 0.2 μL of dNTP mix.
2. The thermal cycling sequence is as follows: the reaction mixture is held at 95°C for 2 min (to activate the Taq polymerase and to denature the genomic DNA)

followed by 40 cycles of 92°C for 10 s, 58°C for 30 s, and 68°C for 30 s. After the final extension step of incubating the reaction mixtures at 68°C for 10 min, they are cooled down and held at 4°C until further use.

3.3. Degradation of Excess PCR Primers and dNTPs

1. Dilute Exo-Sap solution with water in 1:1 ratio.
2. Add 2 µL of the diluted Exo-Sap solution to the PCR product from **Subheading 3.2.**
3. Incubated the mixture at 37°C for 45 min to degrade the excess PCR primers and excess dNTP. Heat inactivate the enzymes were at 80°C for 15 min before the single-base extension reaction.

3.4. Single-Base Extension

1. Make 1 µ*M* stock solution of acycloterminator mix with each of four acycloterminators according to the SNP alleles. For example, prepare a terminator mix containing R110-G, R6G-A, unlabeled acycloterminator C, and T for a G/A SNP.
2. Prepare the TDI cocktail (13 µL) with 1 µL of terminator mix, 0.05 µL of acycloPol polymerase, and 2 µL of 10X TDI buffer.
3. Add the TDI cocktail (13 µL) to the reaction mixtures (7 µL) from the previous step (**Subheading 3.3., step 3**).
4. Add the standard R110 and R6G emission spectra were added to ABI 7700 or 7900 according manufacturer's instructions (*see* **Note 3**).
5. Set up the protocol in ABI 7700 or 7900 sequence detector using R110 as reporter and R6G as Quencher (*see* **Note 4**).
6. The thermal cycling sequence is as follows: incubate the reaction mixture at 95°C for 2 min followed by 40 cycles of 95°C for 10 s, 57°C for 30 s.
7. Collect data at 57°C of each cycle.
8. At the end of the reaction, the mixture is held at 4°C on an ABI 7900 sequence detector.

3.5. Data Analysis

1. Analyze data using the ABI multicomponent option.
2. The results of the multicomponent analysis data can be visualized with ABI software and can then be exported into a text file.
3. The results in text file format can then be imported into any graphing program, such as Excel, SigmaPlot.
4. The original file can be plotted as in the top graph in **Fig. 3**. The quenching is calculated using the intensity of the first cycle minus the intensities of subsequent cycles, the resulting data are plotted as in the bottom graph of **Fig. 3**. The inset is the linear regression of first 8 cycles used to calculate the slope of allele 1 (S1) and the slope of allele 2 (S2).

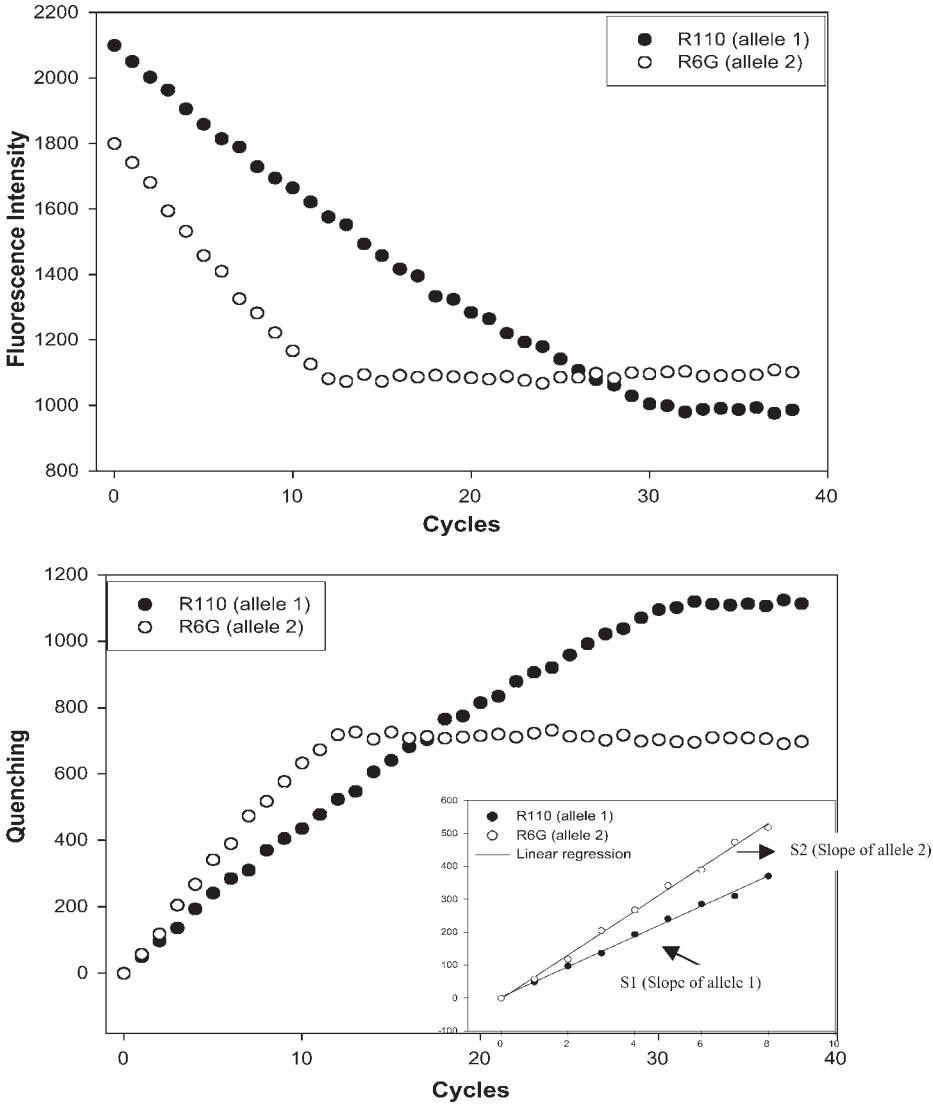


Fig. 3. Data analysis. The top panel is the original data plotted as intensity vs cycles. The bottom panel is the quenching plot, determined as the intensity of first cycle minus the intensity of subsequent cycles. The linear regression of the first eight cycles is shown in the inset and the slopes can be calculated.

5. The allele frequency of allele 1 is determined as $p = 1/[1 + C(S2/S1)]$. C is the correcting factor, which is slope ratio (V_1/V_2) using a heterozygous sample (*see* **Note 5**).

4. Notes

1. The amount of templates from PCR is critical for the FQ-TDI assay; 0.1 pmol PCR templates should be adequate.
2. Because higher quenching will give better results, one can select the sense or antisense primers based on which primer has a G base closer to the 3' end and which primer has more G bases within 10 bases next to 3' end.
3. If R110 and R6G standard emission spectra cannot be added to the ABI software, the standard emission spectra of Fam and Joe can be used for R110 and R6G instead.
4. Although both R110 and R6G are reporters in this assay, the ABI 7700 or 7900 software requires the user to enter a "reporter" and a "quencher" to run the program for data acquisition during thermal cycling. Entering R110 as reporter and R6G as quencher fulfills this requirement and allows the data acquisition to proceed during the primer extension reaction.
5. FQ-TDI works well for SNPs with minor allele frequency between 0.05 and 0.5, and it generates more accurate results for the SNPs with minor allele frequency closer to 0.5. It generally does not work for the SNPs with minor allele frequency less than 0.05, as the slopes of linear regression vary greatly for these SNPs.

References

1. Soto, D., Malmsten, C., Blount, J. L., Muilenburg, D. J., and Caughey, G. H. (2002) Genetic deficiency of human mast cell alpha-tryptase. *Clin Exp Allergy*. **32,** 1000–1006.
2. Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002) DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet*. **3,** 862–871.
3. Xiao M. and Kwok P. Y. (2003) DNA analysis by fluorescence quenching detection. Genome Res. **13,** 932–939.
4. Chen, X., Levine, L., and Kwok, P. Y. (1999) Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res*. **9,** 492–498.
5. Haff, L. A. and Smirnov, I. P. (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res*. **7,** 378–388.
6. Torimura, M., Kurata, S., Yamada, K., et al. (2001) Fluorescence-quenching phenomenon by photo-induced electron transfer between a fluorescent dye and a nucleotide base. *Anal Sci*. **17,** 155–160.
7. Nazarenko, I., Pires, R., Lowe, B., Obaidy, M., and Rashtchian, A. (2002) Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of conjugated dyes. *Nucleic Acids Res*. **30,** 2089–2195.
8. Gardner, A. F. and Jack, W. E. (2002) Acyclic and dideoxy terminator preferences denote divergent sugar recognition by archaeon and Taq DNA polymerases. *Nucleic Acids Res*. **30,** 605–613.
9. Vieux, E. F., Kwok, P. Y., and Miller, R. D. (2002) Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* **Suppl:** 28–32.

10

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

Matthias Wjst and Dirk van den Boom

Summary

Major strengths of mass spectrometry analysis include the accuracy of the detection principle and the automatic data storage, making it a premier choice for high-throughput single-nucleotide polymorphism genotyping. We explain the assay principle in detail and give step-by-step laboratory instructions. Finally, references point toward further use of mass spectrometry analysis for molecular haplotyping, pooled DNA analysis, re-sequencing, and gene expression studies.

Key Words: Matrix-assisted laser desorption mass spectrometry; high-throughput genotyping; haplotype; DNA pooling; gene 1.

1. Introduction

Pharmacogenomic studies rely on genetically determined differences in individuals that are thought to influence treatment response or side effects of a drug. The genetic differences are subtle changes in the nucleotide sequence of the genome. Although there are many repeat regions and small insertions and deletions, the main sources for human genetic variation are single base pair exchanges (single-nucleotide polymorphisms [SNPs]) that occur in functionally important genomic regions.

In diploid species like in humans, SNPs are usually biallelic, although there also are SNPs with three or four alleles. SNPs usually occur every 500 to 1000 bases, leading to an estimate of up to 6 to 10 million SNPs in the human genome. Only a small fraction of these, probably less than 10%, have functional significance, either by influencing the protein amount, the three-dimensional (3D) structure, or functionally relevant amino acids. Many SNPs have been discovered more or less by chance when comparing overlapping sequences from dif-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

ferent individuals. A more systematic approach by comparative sequencing in a defined group of individuals was then started by the SNP consortium. SNP coverage across the human genome turned out not to be random and may be completely absent in highly conserved regions as well as highly abundant in other regions in which genetic diversity is biologically important.

The use of large-scale association studies of genotypes from many individuals participating in a clinical trial is considered to be the most promising method to identify responders and nonresponders to a particular treatment. Relevant SNPs may be situated directly in genes targeted by a specific treatment, for instance, a receptor, as well as in the signaling cascade, even in parallel pathways or in genes involved in the metabolizing pathway of certain drugs.

Intensive research has been conducted worldwide into different assay methods for more than a decade. The technical possibilities to genotype SNPs in individuals have exploded and more than 20 different methods are available today ***(1–3)***. With an estimated count of nearly 10 million SNPs in the human genome, genome-wide SNP genotyping requires high-throughput (HT) methods, but only a few methods are suitable for HT genotyping (usually defined as >5000 genotypes/day/lab) or ultra-HT (>100,000 genotypes/day/lab). Even candidate gene approaches still require substantial genotyping capabilities when studies with sufficient statistical power are to be performed ***(4)***.

1.1. HT Genotyping

Traditional genotyping methods turned out to be unsuitable for HT genotyping. A major requirement for HT genotyping is the automation, from sample preparation to automated readout of the genotype. Another requirement is the availability of sufficient DNA template, the reason why nearly all methods are based on polymerase chain reaction (PCR) amplification. Timing, throughput, and accuracy also are critical. Missing or incorrect genotypes, even in a minor number of samples, may double the time for genotyping. Either individual samples need to be rearrayed in a second step from original plates or repeated from the same source. Average set-up, implementation, and process time for a single assay are therefore important factors to consider. Finally, accuracy is a major demand, as running all assays in duplicate or triplicate would not be cost efficient.

Current methods combine at least one of four different principles of allelic discrimination (hybridization, primer extension, ligation, or restriction) with one of four different detection techniques (chemiluminescence/ fluorescence, fluorescence polarization, resonance energy transfer, and mass spectrometry). Assay formats range from (slab)- gel electrophoresis, plates, particles, fibre arrays, and microchip arrays to semi- and homogenous assays that do not require any further sample separation or purification.

Major strengths of mass spectrometric analysis are the inherent accuracy of this detection principle, the automatic data accumulation and interpretation, and high-throughput capacity ***(5,6)***. The instrumentation comes with slightly higher initial set-up costs compared with other methods, but these amortize very quickly in HT applications. More importantly, the effort required for development and implementation of assays and assay panels is very low. Therefore, mass spectrometry appears to be particularly suitable for fast set-up and analysis of a large number of markers. In addition to the large genotyping capacity, matrix-assisted laser desorption mass spectrometry (MALDI-TOF MS) provides the possibilities of multiplexing and even second-use functions (quantification of allele frequencies, sequencing and even protein analysis), which renders this technology universally applicable.

1.2. Mass Spectrometry

The importance of MS in the field of proteomics and genomics has increased dramatically during the last decade. Although MS has long been a prominent method in analytical chemistry, the analysis of biomolecules appeared to be a problematic task for several reasons.

Generally, in MS an ion source is coupled with a mass analyser equipped with a detection system. The ion source generates gas-phase ions of the molecules of interest. The generation of analyte ions is a prerequisite because mass analyzers usually apply either magnetic or electrical fields for the molecular mass determination. Second, the process of desorption and ionization is a crucial step. It needs to proceed as gently as possible to avoid decomposition of the analyte, and the lack of appropriate methods to produce intact ions of large biomolecules, such as nucleic acids and proteins, initially has hampered the application of MS.

With the introduction of the "soft ionization" methods, electrospray ionization and MALDI. at the end of the 1980s, the accessible mass range for biomolecules was expanded so significantly that both methods now can be seen as cornerstones of modern molecular analysis in proteomics and genomics ***(7)***. This development was rewarded recently with the Nobel price in chemistry 2002 to Fenn and Tanaka. MALDI-TOF MS in particular has significantly impacted the field of nucleic acid analysis. During MALDI, the analyte molecules are mixed with a small molecular weight compound, the matrix. Typically these are small organic molecules with absorption maximum close to the laser wavelength used for subsequent irradiation. The matrix is used in high molar excess over the analyte. The matrix–analyte mixture is then irradiated with a laser beam (lasers emitting in the ultraviolet wavelength or mid-infrared lasers are most common). The irradiation triggers a microexplosion, during which the analyte molecules are co-desorbed into the gas phase with the matrix. The

matrix molecules almost exclusively absorb the laser energy, which allows the generation of intact gas-phase analyte molecules. The most common mass analyzer used with a MALDI ion source is a time-of-flight (TOF) mass analyzer. All ions generated in the desorption process are accelerated to an almost-uniform translational energy by means of an electric field. They then enter a field-free drift region and traverse through this region with a mass-to-charge rate-dependent velocity. The time for traveling through this drift region is recorded and allows determination of the analyte mass.

The use of MS to analyze nucleic acids provides significant advantages. First and foremost, this analytical method determines an inherent physical property of the molecule of interest, the molecular mass. On a principle basis, this provides a higher accuracy than indirect analysis through, for example, fluorescent labels or assessment of gel electrophoretic mobility. The flight time of a molecule is not affected by its 3D structure. Side products sometimes generated in enzymatic reactions usually exhibit a different mass and thus do not lead to misinterpretation of data. Additionally, MALDI-TOF MS provides very high analytical speed. The process proceeds in microseconds and thus provides very fast turnaround times. Mass spectra provide a very simple data format, which lends itself to automated data interpretation without the help of statistical tools. The current rate-limiting step is the laser repetition rate. With current 200-Hz lasers, sample acquisition and real-time data analysis can be completed in 400 ms.

SNP genotyping by MALDI-TOF MS takes advantage of mass differences between allele-specific primer extension products. At present, three related assays are used, the PROBE—primer oligo base extension assay, which was further developed to the MassExtend® assay by SEQUENOM—the PinPoint, and the GOOD assay ***(8)***. A representative scheme is depicted in **Fig. 1**.

The PROBE assay involves a post-PCR primer extension reaction, in which a primer is annealed immediately adjacent to the SNP position and extended allele-specifically to determine the present alleles. The initial implementation of this assay uses a biotinylated PCR primer to enable immobilization of the PCR product to a solid support, such as streptavidin-coated magnetic beads. Immobilization allows for the removal of PCR components and enables the generation of single-stranded template for the subsequent primer extension reaction. Then, a reaction cocktail containing the extension primer, a mix of deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs), along with a thermostable DNA polymerase, is added to the template. The DNA polymerase extends the primer by incorporation of available nucleotides. The reaction terminates if a single dideoxynucleotide is incorporated. The length of allele-specific extension products generated (and hence the corresponding molecular mass) can then be used to identify the possible variants. After purification and

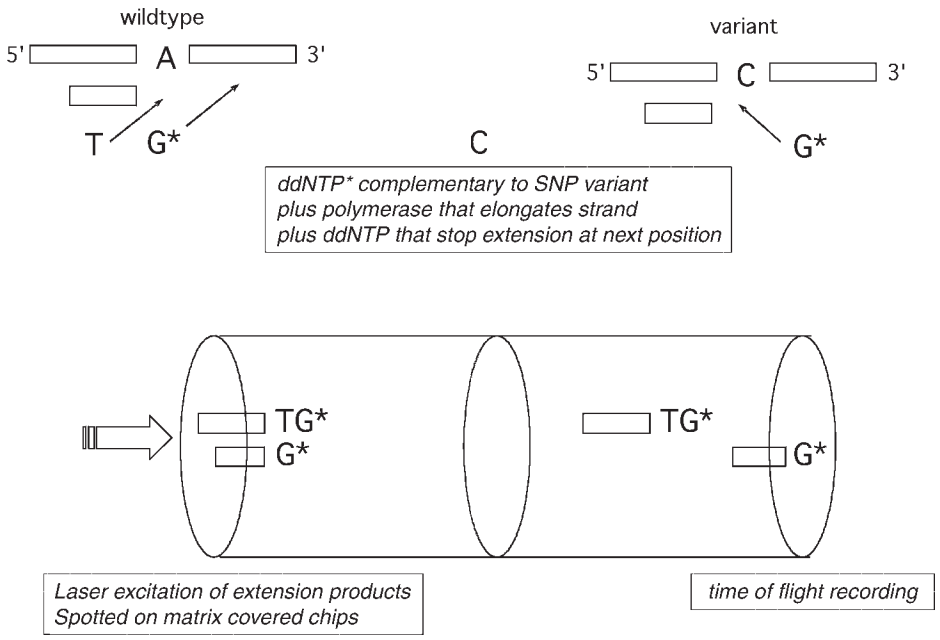


Fig. 1. MALDI-TOF and hME assay principles.

conditioning of the primer extension product, only a few nanoliters of products are transferred onto a prefabricated matrix-loaded chip with a pintool dispenser. Chips can carry as many as 384 samples that are analyzed automatically by MALDI-TOF MS.

To make efficient use of MALDI-TOF MS, the reaction products carrying the genetic information need to be properly conditioned. This usually involves removal of buffer components and especially conditioning of the nucleic acid phosphate backbone (removal of mono- and divalent ions). Initially, solid-phase purification was the preferred method for sample preparation. However, the use of solid-phase protocols is expensive and requires more elaborate sample processing. To minimize these issues, several other strategies were established. The PinPoint assay, for example, uses reversed-phase tip purification, whereas the GOOD assay avoids purification altogether by alkylation of the phosphate backbone enabled through the use of thiolated primer and nucleotides.

The original PROBE assay has been replaced now by a bead-free, homogeneous assay called the MassExtend (hME) assay. Before the primer extension reaction, shrimp alkaline phosphatase (SAP) is added to the PCR product. This dephosphorylates any residual deoxynucleotides that otherwise would inter-

fere with the allele-specific termination. The heat-labile SAP is then easily inactivated. The assay allows a single-tube add-on procedure, in which the addition of ion-exchange resin provides for sample conditioning and is more amenable to automated sample preparation. Sequenom has optimized this protocol for multiplexing up to 15-fold. The following application note summarizes this procedure. A critical step is the use of provided reagents and thermal cycling parameters.

2. Materials

Although all materials may be ordered by individual suppliers, optimized chips and reagents can be ordered on SEQUENOM's web site at www.realsnp.com.

1. MassARRAY Liquid Handler (SEQUENOM, San Diego, CA, cat. no. 11230).
2. MassARRAY Nanodispenser (SEQUENOM, part no. 10024 or no. 10026).
3. MassARRAY Analyzer (SEQUENOM, cat. no. 00450).
4. MassARRAY Typer 3.0.1 or higher (SEQUENOM, cat. no. 11406).
5. MassARRAY Assay Design 2.0 (SEQUENOM, cat. no. 11452).
6. HotStar Taq (QIAGEN).
7. HotStar Taq PCR buffer (QIAGEN).
8. dNTPs (obtain from oligonucleotide supplier).
9. PCR primers (obtain from oligonucleotide supplier).
10. MassEXTEND Starter Kit (SEQUENOM cat. no. 10030).
11. MassEXTEND Mix (SEQUENOM cat. no. 10035-10051).
12. MassEXTEND primers (obtain from oligonucleotide supplier).
13. Thermo Sequenase (SEQUENOM cat. no. 10052).
14. Clean resin (SEQUENOM cat. no. 10053).
15. Clean kit (SEQUENOM cat. no. 11220).
16. SAP (SEQUENOM cat. no. 10002).
17. SpectroCHIP Bioarrays 384-well SpectroCHIP (SEQUENOM cat. no. 00601).
18. Clean Resin Dimple Plate (SEQUENOM cat. no. 11235).

3. Methods

3.1. Assay Design Considerations

For designing high-plexed hME assays, specific primer design software is available that designs PCR and hME primers for each SNP (or insertion/deletion polymorphism) to be investigated. It uses a multiplexing algorithm developed to take full advantage of the available mass range while avoiding overlapping mass signals in the available spectrum range. The program also is designed to consider potential unwanted intra- and inter-primer interactions to avoid misamplification and false-extension products. Before the hME reaction, the genomic DNA is amplified using the PCR (*see* **Notes 1–3**). The use of a 10-mer tag (5'-ACGTTGGATG-3'), referred to as hME-10, on the 5' end of

Table 1
PCR Cocktail

Reagent	Volume	Final concentration
Nanopure water	0.920 µL	NA
Genomic DNA (2 ng/µL)	1.000 µL	2 ng/rxn
HotStar Taq® PCR buffer[a] containing 15 m*M* $MgCl_2$ (10X)	0.625 µL	1.25X/1.875 m*M* $MgCl_2$
Fresh dNTPs (25 m*M*)[b]	0.100 µL	500 µ*M* each
Forward PCR primers (500 n*M* each)[c]	1.000 µL	100 n*M* each
Reverse PCR primers (500 n*M* each)[c]	1.000 µL	100 n*M* each
$MgCl_2$ (25 m*M*)	0.325 µL	1.625 m*M*[d]
HotStar Taq (5U/µL) QIAGEN Inc.	0.030 µL	0.15 U/rxn
Total	5.000 µL	

[a]The PCR buffer concentration should not exceed 1.25X. Higher salt concentrations have negative effects at the hME level.
[b]Maximum of 5 freeze/thaws.
[c]Containing a 10-mer tag: hME-10 (5'-ACGTTGGATG-3').
[d]3.5 m*M* $MgCl_2$ total. Do not use Q solution. It has negative effects on MALDI-TOF MS analysis.

each PCR primer provides significant improvement in overall hME performance. The tags increase the masses of unused PCR primers so that they fall outside the mass range of analytical peaks and help to balance amplification.

3.2. Polymerase Chain Reaction

To prepare and process the PCR, perform the following steps:

1. Prepare a PCR cocktail as described in the table (volumes are provided on a per-well basis)
2. Cycle the PCR as follows in a standard thermal cycler:

0.3 U of SAP provided in a 2-µL volume of enzyme is then added to each PCR to dephosphorylate unincorporated dNTPs from the amplification reaction.

3.3. Adjusting Primer Amount

The peaks in the mass spectrum for a multiplexed reaction may not have comparable heights. Variations in peak height may stem from 1) inconsistent oligonucleotide quality, 2) inconsistent oligonucleotide concentration, and 3) different desorption/ionization behavior in MALDI. For best multiplexing results, the concentrations of hME primers should be adjusted to even out peak heights (intensities) in the mass spectrum. This adjustment must be done before preparing the hME reaction cocktail and processing the hME reaction.

Table 2
PCR Conditions

Cycles	Condition
1	95°C for 15 min
45	95°C for 20 s
45	56°C for 30 s
45	72°C for 1 min
1	72°C for 3 min
1	4°C hold

The following steps need to be performed to adjust primer mixes:

1. For each multiplex, prepare a mixture of the required primers. The final concentration of each primer in the primer mix should be 9 μ*M*. Consider how much primer mix you will need so that this step has to be performed only once for the assay set-up. Each single reaction (i.e., a single well in a 384-well microplate) requires 1 μL of primer mix.
2. Pipet 1 μL of the primer mix into a well of a microplate and add 24 μL of nanopure water to obtain a 360 n*M* dilution of the primer mix (referred to as a primer mix sample).
3. Repeat **steps 1** and **2** for each multiplex to generate a microplate containing primer mix samples for all of the multiplexes.
4. Add 3 mg of resin to each well of the microtiter plate using the dimple plate.
5. Dispense the primer mix samples to a precoated chip using standard dispensing conditions for hME reaction products.
6. Acquire spectra from using MassARRAY™ typer software 3.0.1 or higher. Use the assay definitions (in Typer) for the actual multiplexes. Each well on the SpectroCHIP® bioarray will yield no-calls because there is no analyte, only unextended primers. A peak should appear at the expected mass for each primer in the mix. A missing peak generally indicates poor primer quality or a primer missing from the mix. An unexpected peak generally indicates poor primer quality or the addition of an unnecessary primer to the mix.
7. Check whether the primer peaks in each mass spectrum have comparable heights (*see* **Note 4**). If all peaks are at least 50% the height of the highest peak, they are acceptable. If any peak is less than 50% the height of the highest peak, add more of that primer, for example, add the deficit in percent from the highest peak as percent of the initial volume. A corresponding report function is provided within the supplied genotyping software (*see* **Notes 5– 7**).

3.4. hME Reaction, Desalting, and Dispensing

Once the hME primer mixes have been adjusted, the hME reaction cocktail is prepared, added to the SAP-treated PCR product, and thermocycled.

Table 3
Extension Reaction Cocktail (Per Reaction Well)

Reagent	Volume	Final concentration
Nanopure water (high-performance liquid chromatography grade)	0.760 µL	NA
PCR product, including buffer and d/ddNTPs	0.200 µL	50 µ*M* each d/ddNTP
Adjusted primer mix (~9 µ*M* each)[a]	1.000 µL	1.25X/1.875 m*M* $MgCl_2$
Thermo Sequenase (32 U/µL)	0.040 µL	1.25 U/rxn

[a]Note that the primers in an adjusted mix may not be at 9 µ*M* each. Each starts out at 9 µ*M*; however, the addition of extra amounts of some primers to adjust the mix will change the concentrations.

Then, add this 2 µL of the hME reaction cocktail to the SAP-treated PCR products. Cycle the hME reaction as follows.

Table 4
PCR Conditions

Cycles	Condition
1	94°C for 2 min
75	94°C for 5 s
75	52°C for 5 s
75	72°C for 5 s
1	4°C hold

Dilute with 16 µL and add 6 mg of clean resin to the hME reaction products for conditioning (*see* **Note 3**). Then incubate for 5 min at room temperature and keep the resin particles in suspension during incubation. Spin the reaction vessel before the next step. Using a nanodispenser, 15 nL of the reaction product are then dropped onto a precoated 384-well SpectroCHIP.

3.5. Desorption and Spectral Analysis, Assignment of Genotypes

Analysis of chip-transferred samples proceeds in a linear, delayed extraction TOF mass spectrometer. Mass spectra are acquired in positive ion mode (all positively charged molecular ions are accelerated). The chips are introduced into the ion source, and high-vacuum conditions are applied. Image processing aligns the laser position automatically to the chip element raster for

fully automated scanning of each chip position. Each matrix crystal is addressed individually and irradiated with a 337-nm laser pulse of 1-ns duration. The irradiation results in a plume of volatized matrix and analyte. During gas phase, charge-transfer processes generate matrix and analyte ions, which are accelerated in an electric field. By traveling through a field-free region of approx 1 m in length, their velocity is inversely proportional to their mass-to-charge ratio. The resulting time-resolved mass spectrum is then translated into mass spectrum by comparison with known calibrants. Usually, 15 single laser shots are accumulated and averaged into a single spectrum. This average spectrum is then further processed and analyzed using dedicated software (SPECTRO TYPER RT, SEQUENOM) that performs baseline correction, peak identification and quality assessments. The determination of corresponding genotypes occurs real time during data acquisition and is usually completed within one second processing time (transit time of laser, laser irradiation, spectra accumulation and analysis). If the mass spectrum is not of sufficient quality, the software will automatically reacquire new data points from the same chip position before it moves to the next chip position. This provides real-time control of data quality and increases accuracy as well as call rates.

3.6. Other MALDI Applications

The focus of this chapter has been genotyping of SNPs using primer extension methods and MALDI-TOF MS. Within recent years, the portfolio of applications using MALDI-TOF MS as a detection platform has expanded significantly. A majority of these new applications not only rely on the accuracy provided by MS for qualitative analysis of nucleic acids, but they also have established measures for quantitative analysis of nucleic acids. Recent publications describe the use MALDI-TOF MS for relative quantitation of genetic information in DNA pools and sample mixtures ***(9–13)***; re-sequencing methods, which allow the rapid discovery of SNPs, the screening for mutations or signature sequence based identification of organisms such as pathogens ***(14–16)***; and also relative and absolute quantitation in gene expression ***(17)***. A further interesting application is M1-PCR for haplotyping ***(18)***. Here, multiplex PCR performed on single DNA molecules generated by dilution is combined with the specificity of mass spectrometry read-outs to generate up to 25 kB haplotypes. Recent reviews summarize these developments ***(19,20)***.

4. Notes

In addition to the above procedures it seems to be worthwhile to also consider the following points.

1. PCRs for the MassEXTEND reaction are usually performed in low volumes (5 µL). It is important that the TE concentration in the genomic DNA does not inhibit the

Table 5
SNPs and Suitable Termination Mixes

SNP	Termination mix
A/C	dATP/ddCTP/ddGTP/ddTTP
A/G	dGTP/ddATP/ddCTP/ddTTP
A/T	dATP/ddCTP/ddGTP/ddTTP
C/G	dGTP/ddATP/ddCTP/ddTTP
	dCTP/ddATP/ ddGTP/ ddTTP
C/T	dTTP/ddATP/ddCTP/ddGTP
G/T	dGTP/ddATP/ddCTP/ddTTP
Ins/ dels	Dependent on sequence context

amplification. Make sure that the genomic DNA does not contain more than 0.25X TE buffer.

2. The matrix/crystallization process is sensitive to detergents. PCR additives, such as Q solution (provided with HotStarTaq), may disturb the crystallization process and reduce the data quality and thus should be avoided.
3. Desalting the MassEXTEND products with CLEAN resin is a crucial step with strong impact on the data quality. It is important that the resin particles stay in suspension during the 5-min incubation step and do not settle. A rotation where plates are turned upside down usually provides best performance. Increased incubation temperature is not recommended.
4. Oligonucleotides of poor quality (increased amount of synthesis failure products or strong depurination signals) will lead to poor genotyping performance and may interfere with correct genotype assignment. Make sure during the primer amount adjustment that each primer generates only the desired mass signal. Preferably, order primers from oligonucleotide manufacturers using MALDI-TOF MS for synthesis quality control.
5. When designing genotyping assays manually, do not use termination mixes containing all four dideoxynucleotides. Mass differences between alleles would then be as little as 9 mol mass (ddATP/ddTTP mass difference). This can be challenging to discriminate and can lead to wrong genotype assignments. Additionally, the mass difference between ddA/ddC and ddT/ddG falls close to the mass of sodium adducts (22 mol mass), potentially leading to misinterpretation of mass signals.
6. When designing assays manually, check PCR primers for multiple binding to the genome and for formation of primer dimers and hairpins to avoid misamplification. Check self-designed MassEXTEND primer for hairpin formation to avoid self-extension. A commercial software package is available from SEQUENOM.
7. Occasionally DNA polymerase pausing has been observed when the template exhibits strong secondary structure. This leads to prematurely terminated exten-

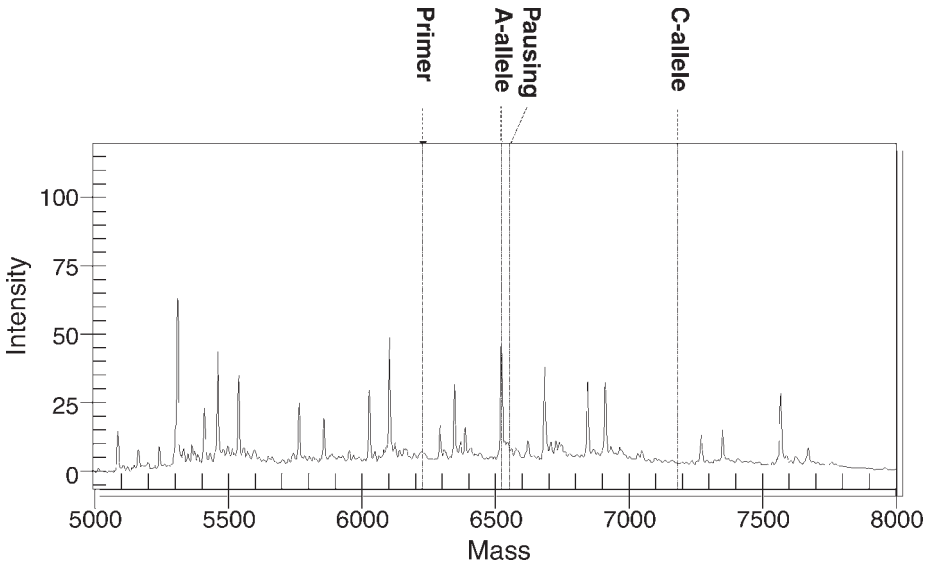


Fig. 2. 12-plexed MassEXTEND reaction results. To exemplify the concept, the assay definitions for one of the 12 assays (primer, potential pausing signals, and predicted alleles) are indicated with dotted lines. In the depicted case, the individual is homozygous A. The MassEXTEND primer is fully extended. Each mass signal provides a unique identifier for the presence of an allele in a specific assay.

sion products, which can confound the analysis if termination mixes are not selected carefully (note that an extension primer elongated either with one ddGTP or dATP will have the same molecular mass). The following table provides a list of suitable termination mixes for biallelic SNPs, which prevent mass signal coincidence of pausing artefacts and real termination events:

References

1. Syvänen, A. C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* **2,** 930–938.
2. Kwok, P. Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.* **2,** 235–258.
3. Tsuchihashi, Z. and Dracopoli, N. C. (2002) Progress in high throughput SNP genotyping methods. *Pharmacogen J.* **2,** 103–110.
4. Little, J., Bradley, L., Bray, M. S., et al. (2002) Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am. J. Epidemiol.* **156,** 300–310.
5. Ross, P., Hall, L., Sminov, I., and Haff, L. (1998) High level multiplex genotyping by MALDI TOF mass spectrometry. *Nat. Biotechnol.* **16,** 1347–1351.

6. Kim, S., Ruparel, H. D., Gilliam, T. C., and Ju, J. (2003) Digital genotyping using molecular affinity and mass spectrometry. *Nat. Rev. Genet.* **4,** 1001–1009.
7. Berkenkamp, S., Kirpekar, F., and Hillenkamp, F. (1998) Infrared mass spectrometry of large nucleic acids. *Science* **281,** 260–262.
8. Tost, J. and Gut, I. G. (2002) Genotyping single nucleotide polymorphisms by mass spectrometry. *Mass Spectrom. Rev.s* **21,** 388–418.
9. Ross, P., Hall, L., and Haff, L. A. Quantitative approach to single nucleotide polymorphism analysis using MALDI TOF mass spectrometry. *BioTechniques* **29,** 620–629.
10. Bansal, A., van den Boom, D., Kammerer, S., et al. (2002) Association testing by DNA pooling: an effective initial screen. *Proc. Natl. Acad. Sci. USA* **99,** 16,871–16,874.
11. Werner, M., Sych, M., Herbon, N., Illig, T., König, I., and Wjst, M. (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. *Hum. Mutat.* **20,** 57–64.
12. Mohlke, K. L., Erdos, M. R., Scott, L. J., et al. (2002) High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl. Acad. Sci. USA* **99,** 16,928–16,933.
13. Herbon, N., Werner, M., Braig, C., et al. (2003) High-resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases. *Genomics* **81,** 510–518.
14. Stanssens, P., Zabeau, M., Meersseman, G., et al. (2004) High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. *Genome Res.* **14,** 126–133.
15. Elso, C., Toohey, B., Reid, G. E., Poetter, K., Simpson, R. J., and Foote, S. J. (2002) Mutation detection using mass spectrometric separation of tiny oligonucleotide fragments. *Genome Res.* **12,** 1428–1433.
16. Lefmann, M., Honisch, C., Bocker, S., et al. (2004) Novel mass spectrometry-based tool for genotypic identification of mycobacteria. *J. Clin. Microbiol.* **42,** 339–346.
17. Ding, C. and Cantor, C. R. (2003) A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight, M. S. *Proc. Natl. Acad. Sci. USA* **100,** 3059–3064.
18. Ding, C. and Cantor, C. R. (2003) Direct molecular haplotyping of long-range genomic DNA with M1-PCR. *Proc. Natl. Acad. Sci. USA* **100,** 7449–7453.
19. Jurinke, C., Oeth, P., and van Den Boom, D. (2004) Maldi-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. *Mol. Biotechnol.* **26,** 147–164.
20. Jurinke, C., van den Boom, D., Cantor, C. R., and Köster, H. (2002) The use of MassARRAY technology for high throughput genotyping. *Adv. Biochem. Eng. Biotechnol.* **77,** 57–74.

11

Fluorescence-Based Fragment Size Analysis

Peter Imle

Summary

The successful application of capillary electrophoresis technology to the genotyping of various types of polymorphisms has been well documented. The flexibility and automation of the Applied Biosystems 3100 Genetic Analyzer make it an excellent capillary electrophoresis platform for the generation of high quality genotype data. These data are readily applied to pharmacogenomic investigations of various types. Included in this chapter is a protocol for the generation of genotype data using minimal template deoxyribonucleic acid and maximizing the automation of both data analysis and genotype assignment through the use of the Applied Biosystems GeneMapper 3.0 software package.

Key Words: Fragment size analysis; capillary electrophoresis; genotyping; polymorphism; polymerase slippage.

1. Introduction

The evolution of deoxyribonucleic acid (DNA) fragment size analysis from slab gels using radiolabeled or carcinogenic stains to capillary electrophoresis and fluorescence-based detection has provided for substantial increases in automation, accuracy, and precision *(1,2)*. These advantages provide a cost effective, high-throughput screening method that can be utilized in a variety of pharmacogenomic investigations. The flexibility of this technique makes it a useful tool for the analysis of polymorphisms that cause an in vivo change in the size of a DNA fragment. This allows for the screening of a variety of mutational events, such as expansion or contraction of tandem repeats, insertions and deletions of varying size, or with additional manipulation of single nucleotide polymorphisms *(2,3)*. The primary steps in a fragment size analysis protocol are the generation of the fragments containing the polymorphism of interest, the separation and detection of these fragments by capillary electro-

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

phoresis, determination of fragment size, and assignment of alleles. Fluorescence-based capillary electrophoresis requires the attachment of synthetic compounds designed to emit specific wavelengths of light upon excitation. Most high-throughput capillary electrophoresis systems use laser-induced fluorescence for excitation of these compounds and an on-line CCD camera for detection of the emitted light ***(4)***. The attachment of the fluorescent compounds can be done before the detection of the polymorphism by synthesizing one member of a forward and reverse primer pair with the fluorescent label attached to the 5' end. Use of these primers in a polymerase chain reaction (PCR) will generate amplicons that have incorporated the fluorescent label and will be detectable by capillary electrophoresis systems that are equipped with fluorescence detection capabilities.

2. Materials

1. DNA samples.
2. 96-well thin wall plates suitable for thermal cycling (MJ Research, cat. no. MLL-9631).
3. True Allele PCR Premix (Applied Biosystems, Foster City, CA, cat. no. 4013061).
4. Primer stocks: 2.5 µ*M* stocks of each labeled forward and unlabeled reverse primer (*see* **Notes 1** and **2**).
5. DNA Thermal Cycler (MJ Research) PTC-200 or PTC-225 can be used.
6. Redistilled formamide (Invitrogen, cat. no. 15515-026).
7. Resin (Bio-Rad, cat. no. AG501X8).
8. 96-well optical plates (Applied Biosystems, cat. no. N801-0560).
9. ROX400HD size standard (Applied Biosystems, cat. no. 402985).
10. 3100 Genetic Analyzer (Applied Biosystems, cat. no. 3100-01).
11. GeneMapper 3.0 analysis software (Applied Biosystems, cat. no. 4319372; *see* **Note 3**).

3. Methods

3.1. PCR Amplification

1. Assemble a 15-µL reaction containing 9 µL of TrueAllele PCR Premix, 2 µL of a 2.5 µ*M* stock of each forward and reverse primer, and 2 µL of DNA (for a total of 5 ng; *see* **Notes 4** and **5**).

Reagent	Volume	Final concentration
True Allele PCR Premix	9 µL	1X
Labeled Forward Primer (2.5 µ*M*)	2 µL	0.33 µ*M*
Unlabeled Reverse Primer (2.5 µ*M*)	2 µL	0.33 µ*M*
DNA (2.5 ng/µL)	2 µL	5 ng

2. PCR conditions
   a. 95°C for 10 min (hot start heat activation of the polymerase).
   b. 30 cycles: 95°C for 1 min (Denature), 60°C for 1 min (Annealing), and 72°C for 1 min (extension)
   c. 72°C for 10 min (final extension of products).
   d. Hold products at 4°C until ready for processing.

3.2. Preparing PCR Products for Injection and Electrophoresis

1. Deionize Formamide: add 5 mL of redistilled formamide to 1.0 g of resin and let sit for 15 min to deionize.
2. To prepare sample for injection add 9.5 µL of deionized Formamide to 96-well optical plate, add 0.5µL of ROX400HD standard, and 1 µL of PCR product.
3. Denature mixture at 95°C for 5 min
4. Chill plate at 4°C for 2 min.
5. Briefly centrifuge sample to collect sample at the bottom of the well and to remove air bubbles.
6. Select run module for injection and electrophoresis conditions.

3.3. Data Analysis

The data analysis process begins with the identification of the primary peaks(s) from the sample electropherogram. Next is the calculation of the size, in base pairs, of the identified peak(s) followed by the assignment of alleles to complete the genotype assignment for each sample. GeneMapper 3.0 (Applied Biosystems Incorporated, Foster City, CA) is an integrated software package that partially automates these steps with minimal user intervention ***(5)***.

Identification of the primary peak(s) from within the electropherogram requires the filtering of a variety of confounding information. Genemapper 3.0 software uses algorithms that are designed to look for common problems such as high background fluorescence, minimum and maximum peak height, the presence of "stutter" peaks that result from replication errors generated by polymerase slippage ***(6)***, and the presence of nontemplate-mediated adenosine addition generating fragments one base pair larger than expected ***(7)***. Problems related to peak height can often be alleviated by changing either the amount of PCR product added to the injection mixture or varying the time of the injection cycle for the electrophoresis (*see* **Note 6**). Polymerase slippage is a common problem in the generation of PCR fragments containing highly repetitive regions. The shorter the repeat unit, the more prevalent the replication errors, as is clearly shown in **Fig. 1**. Generation of stutter peaks in 3 and 4 bp repeat fragments is rarely a problem. The most common solution for minimizing the generation of these fragments is to increase the annealing temperature to increase the stringency of the PCR cycling.

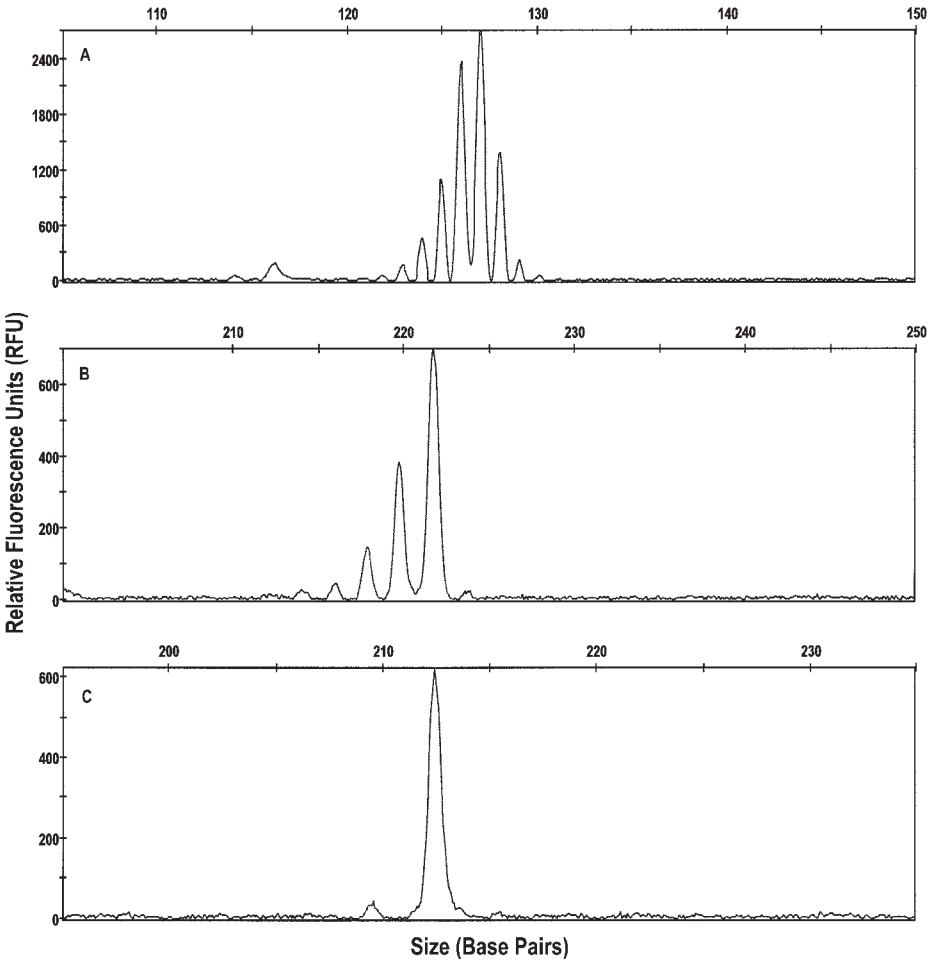


Fig. 1. Results of polymerase slippage in 1-, 2-, and 3-bp repeats. **(A)** shows the number of fragment species generated from a sample homozygous for a 26-bp pair poly-A repeat. The generation of as many as seven different detectable fragments after both PCR optimization and primer "pig-tailing" is common. The number of species generated by polymerase slippage in the 2-bp repeat **(B)** and a 3-bp repeat **(C)** are greatly reduced.

Taq polymerase also is known to commonly add an additional Adenosine to the 3' end of an amplicon that is not present in the template. The occurrence of this +A peak can lead to the generation of multiple confounding peaks in all types of fragments, however, the most problematic cases develop in shorter repeat units. The electropherograms in **Fig. 2** show the extent to which this

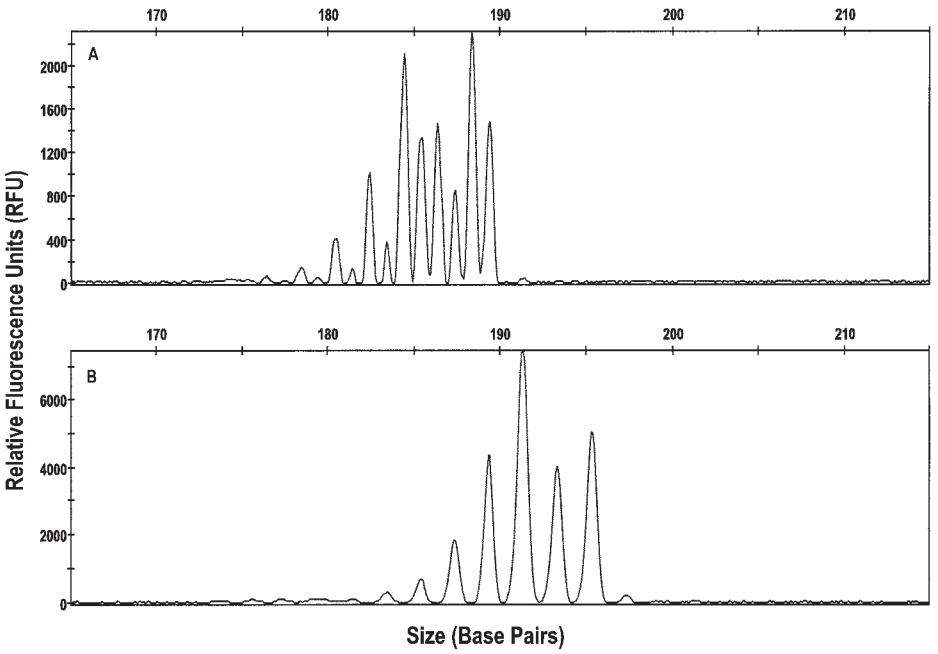


Fig. 2. The effects of "pig-tailing" the amplification primers to maximize the Taq adenosine addition. The addition of an adenosine to the 3' end of amplicons that is not matched in the amplification template is shown to greatly increase the number of observed fragment species even after the optimization of PCR conditions in a sample heterozygous for a 2-bp repeat **(A)**. The addition of the GTTTCT "pig-tail" to the 5' end of the reverse primer greatly increases the affinity of this addition reducing the number of observed fragments **(B)**. Note the increase in overall fragment size of seven base pairs, which represents the 6-bp primer addition and the added adenosine.

addition can confound allelic designation. It has been previously shown that different sequence motifs have varying affinities for the nontemplate-mediated adenosine addition *(7)*. This work also suggested that the identification of sequence motifs that encouraged this +A addition were more effective than those that attempted to reduce its occurrence. Consequently, the synthesis of primers that contain a GTTTCT "pig-tail" of unique sequence on the 5' end can raise the affinity for the addition of adenosine. This increased affinity causes a higher proportion of the generated fragments to contain fragments of a single species, the known length plus the additional adenosine base, greatly simplifying the data analysis.

Once the primary peak(s) in an electropherogram have been identified they are assigned a size in base pair by the GeneMapper 3.0 software. The sizing

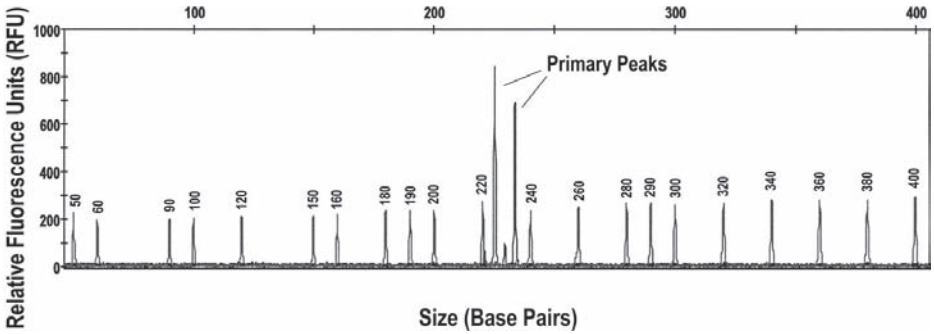


Fig. 3. The use of internal size standards for accurate sizing. Size standards that are run concurrently with each of the sample allows for the accurate and consistent sizing of the primary fragment peaks. This figure shows the sizes for each of the known fragments from the standard superimposed on the unknown fragments. The primary peaks of the unknown fragments are labeled. The GeneMapper 3.0 sizing algorithm uses this comparison for its assignment of fragment size.

peak sizing is done by the comparison of the detected peaks with those identified in the internal size standard, ROX400HD, that is run concurrently with the unknown sample. This size standard is run in a different color detection channel, so there is no peak overlap with the unknown sample. **Figure 3** shows an overlay of the two different channels representing the standard peaks, labeled with their known sizes, and the unknown primary peaks identified by the GeneMapper 3.0 software. The peaks present in the size standard provide the algorithm with a reference for the sizing of the unknown peaks. This concurrent running of the size standard helps maintain a consistency in sizing and allele calling from run to run and allows comparison of data generated over long periods of time.

4. Notes

1. Synthesis of primers must take into account the addition of the proper synthetic fluorescent compound attached to the 5' end of the forward primer. This synthesis results in the accumulation of unincorporated fluor that must be removed by purification. Successful removal of both unincorporated fluor and incomplete synthesis products can be done by either high-performance liquid chromatography or reverse-phase column purification.
2. Selection of a fluorescent compound is dictated by the detection capabilities of the capillary electrophoresis instrument. The Applied Biosystems 3100 Genetic Analyzer has a variety of supported detection colors, as described in the following reference guide. Most color combinations include multiple detection channels for the detection of both samples and concurrently run standards used by the sizing algorithm.

3. An alternative to the commercial GeneMapper 3.0 genotyping software package is the freeware GeneScanView package, developed by Davide Campagna at CRIBI, University of Padova, Italy. This software is capable of reading fragment size analysis files generated by the Applied Biosystems 3100 Genetic Analyzer and generates an electropherogram for visualization and exact sizing of the data.
4. DNA template amounts from 5 to 50 ng from a variety of purification methods have been successfully used in this Fragment Size Analysis protocol. Of primary importance is the uniformity of the concentrations across all samples. This will aid in the uniformity of sample injection, providing higher quality data for downstream analysis.
5. Because of the limited amount of sample that is used in the injection procedure, smaller PCRs can be used to reduce reagent use. PCRs as small as 5 μL have been used successfully in this protocol if special care is taken to minimize evaporation of the samples during amplification. The use of 384-well plates and heat activated plate seals can help minimize this problem.
6. Variation in observed fluorescence can be controlled by changing the amount of PCR product that is added to the deionized formamide dilution reagent. Additionally, changes in the duration of the sample injection step of the capillary electrophoresis run module can have a stabilizing effect on the amount of sample that is loaded into the capillary. Reducing the injection time has a more profound effect on those reactions that contain excess amounts of sample than those having limited amounts. This causes a greater reduction in signal strength for those samples that may be exceeding the detection capabilities of the instrument than those exhibiting minimal signal strength. Increases in injection time can assist in recovering samples that amplify poorly. However, this often increases the amount of background fluorescence that is observed as a result of the injection of unwanted material, such as carryover PCR reagents and remnants of DNA extraction materials, which can complicate the downstream data analysis and should be used cautiously.

References

1. Mitchelson, K. R. and Cheng, J., (ed.) (2001) *Capillary Electrophoresis of Nucleic Acids, Vol I*. Humana, Totowa, NJ.
2. Mitchelson, K. R. and Cheng, J., (ed.) (2001) *Capillary Electrophoresis of Nucleic Acids, Vol II*. Humana, Totowa, NJ.
3. Huang, X. H., Salomake, A., Malin, R., Koivula, T., Jokela, H., and Lehtimaki, T. (1997) Rapid identification of angiotensin-converting enzyme genotypes by capillary electrophoresis. **43,** 2195–2196.
4. Mitchelson, K. R. (2001) The application of capillary electrophoresis for DNA polymorphism analysis, in *Capillary Electrophoresis of Nucleic Acids, Vol I*. (Mitchelson, K. R. and Cheng, J., eds.), Humana, Totowa, NJ, pp. 3–26.
5. Applied Biosystems, Inc. (2002) *ABI Prism GeneMapper Software Version 3.0: User's Manual*. Foster City, CA (no. 4335526 Rev. B)

6. Shinde, D., Lai, Y., Sun, F., and Arnheim, N. (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. *Nucleic Acids Res*. **31,** 974–980.
7. Brownstein, M. J., Carpten, J. D., and Smith, J. R. (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* **20,** 1004–1006, 1008–1010.

12

Single-Nucleotide Polymorphism Genotyping in DNA Pools

Ian Craig, Emma Meaburn, Lee Butcher, Linzy Hill, and Robert Plomin

Summary

To undertake partial, or complete, genome screens by association-based methodology for quantitative trait loci, multiple individuals have to be screened for large numbers of genetic markers. Consequently, much recent interest has focused on methods enabling accurate allele quantification in pooled deoxyribonucleic acid (DNA) samples. Microsatellites were the favored markers in initial studies, but the extraordinary wealth of data concerning single-nucleotide polymorphisms (SNPs) has turned attention to the quantification of SNP alleles in pools. All such approaches require accurate estimation of DNA concentrations, followed by the preparation of replicate pools, their validation, and application of procedures for determining allele frequencies. This chapter describes the important steps in preparing pools and surveys a variety of techniques that have been proposed for SNP detection. Finally, we describe the application of a generic approach using pooled DNA for detection of allele frequency differences between case and control populations based on primer extension protocols and outline a strategy for estimating SNP allele frequencies employing microarrays.

Key Words: Pooling DNA samples; DNA quantification; single-nucleotide polymorphisms (SNPs); micro-arrays.

1. Introduction

1.1. General Considerations

The last decade has seen extraordinary success in the use of human linkage maps for identifying the genetic variants responsible for Mendelian, single-gene disorders. The virtues of linkage analysis are that it can be adapted to examine the co-segregation of a marker and a disease-conferring locus under a variety of proposed models that incorporate different inheritance patterns, penetrance values, and so on. A particular advantage is that linkage uses relatively

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

few markers to scan the entire genome, typically using approx 400 evenly spaced microsatellites. Its range can extend over long genetic distances (up to 20 c*M*). For these reasons, linkage analysis is ideally suited for a systematic screen of the genome for disease-predisposing loci. Linkage analysis, however, has relatively low power to detect genes of small effect, and for this reason it has been relatively unsuccessful in applications to identify genes implicated in common disorders. Here, the expectation is that manifestation of the phenotype will depend on an underlying quantitative distribution of factors, both genetic and environmental. Many genes may each contribute only slightly to the overall predisposition and for this reason are referred to as quantitative trait loci (QTLs). Indeed, the challenge is to be able to detect QTLs that contribute as little as 1% to the genetic variance ***(1)***. Detection of such loci is essentially beyond the power of linkage analysis.

Association studies, in contrast, have the potential to detect genes of small effect. The sacrifice made to achieve this, however, is that markers must be the risk variant itself, or very tightly linked to the risk variant. For this reason, association analysis has until recently been used mainly for the study of polymorphisms within, or around, candidate genes that are involved in one of the pathways hypothesised to be relevant to disease aetiology. The possibility of using association analysis in a systematic screen of the entire genome was raised by Risch and Merikangas ***(2)***, who showed that, even if all the functional variants in the genome were examined and an appropriate Bonferroni correction were made for all these tests, association analysis would still have far greater power than linkage for detecting genes of minor effect.

The number of markers required for such an endeavour is likely to be large, possibly even up to hundreds of thousands ***(3)***, because successful detection of an association requires that the marker and QTL must be in linkage disequilibrium (LD) and this rapidly declines as a function of increasing genetic distance (e.g., **ref. *4***). Patterns of LD across the genome are highly variable, and to achieve reasonable coverage, intermarker distances of the order of 10 kb are thought to be necessary.

Thus, it seems likely that genome-wide association studies will require genotyping tens of thousands of markers. Furthermore, it is generally accepted that complex traits are influenced by many genes of varying, but small, effect size. Power to detect QTLs of 1% effect size will require thousands of individuals. There is, therefore, an urgent need to develop efficient high-throughput methodologies for detecting such markers. One way to reduce costs dramatically is to perform genotyping not on individual deoxyribonucleic acid (DNA) samples, but on pools made up of DNA from multiple individuals. For example, the allele frequencies in a sample of 500 cases and 500 controls can be measured from two pooled samples, rather than 1000 individual samples,

reducing the genotyping costs by a factor of 500. Developments of this nature, coupled with high-throughput SNP genotyping methodologies, suggest that genome-wide association studies may soon become a feasible proposition (e.g., **refs. *5*,*6***).

Unfortunately, life is not so simple. There are problems in the use of pooled DNA for association analysis. For quantitative traits, pooling will result in loss of information concerning the variation among individuals of the same pool and in particular, it is difficult to examine multilocus haplotype effects or interactions, although some useful information can be extracted ***(7)***. Furthermore, allele frequency estimation from pooled DNA is subject to both bias and random measurement errors; however, careful attention to experimental design can minimize such difficulties ***(5*,*8)***. Nevertheless, little doubt exists that for laboratories equipped with conventional apparatus and resources, pooling provides the only feasible approach to large-scale, genome-wide association studies. In this chapter, we describe methodology that is designed to minimize problems of pool construction and provide sample protocols for two approaches designed to estimate SNP allele frequencies in pools.

1.2. SNP Detection and Allele Frequency Determination

SNPs generally reflect the existence of two alleles at appreciable frequency at a given nucleotide site. Although it is theoretically possible to have more than two alleles in the population, the following discussion assumes a simple biallelic system. The concept behind pooling subassumes that the constructed pool will represent the molecular equal equivalents of DNA from all members contributing to the pool. This requires two conditions to be met. Firstly, the DNA sampling and quantitation techniques are sufficiently sensitive and accurate to enable equimolar amounts of DNA from each individual to be combined. Second, the SNP genotyping assay works equally efficiently on each individual sample, so that each contributes proportionately to the final allele balance. At the technical level, the former is easier to achieve and monitor than the latter (*see* **Subheading 1.4.**). The latter depends upon the DNA samples being of equivalent integrity with regard to fragmentation and/or presence of interfering contaminants. One way to determine the equivalent integrity of the DNA samples is to employ TaqMan™, or similar real-time polymerase chain reaction (PCR) approaches, to assess the amplification efficiency and, on the basis of the results, adjustments made so that each individual provides a similar number of template equivalents. Finally, because the detection efficiency may differ for the two alleles of a SNP, the application of a correction factor may be necessary to generate better estimates of the "true" frequencies. This factor generally referred to as "*K*," can be derived by examining the relative signal intensities obtained for the two alleles in known heterozygotes ***(9)***.

Historically, restriction enzyme cleavage was the first technology applied to the detection of SNPs (i.e., restriction fragment length polymorphisms) and it has been applied to DNA pools ***(10)***; however, the problem of partial digestion and the relatively restricted number of potential target sites severely limits its applicability. There are now, however, a variety of technologies that have been developed for the estimation of SNP allele frequencies (*see* **refs. *5*,*11*** for reviews), many of which have been adapted for profiling of pooled DNA. **Table 1** lists the strategies, applications and references for those most frequently employed.

In many approaches, laser excitation of the fluorescently tagged products results in peaks representing the two alleles, whose heights in pooled samples are proportional to the allele frequencies. For some procedures, it is claimed that frequency estimates approaching the level of individual genotyping errors (i.e., 1–2%) can be achieved (*see* **Table 1** ***[12–33]***). Other strategies, such as Pyrosequencing™, and those based on denaturing high-performance liquid chromatography, also can be modified for pooling analysis. Even higher throughput for SNP genotyping can be achieved by microarray and/or mass spectrometry technologies and, again, these techniques are being adapted for estimation of allele frequencies in pools.

Given the wide variety of approaches now available, further practical considerations in this chapter will be confined to SNP detection using microsequencing and to novel microarray-based technologies. Microsequencing has proved to be a robust technology, and kits are commercially available for ABI, Applied Biosystems sequencers, MegaBACE (Amersham Biosciences), and some other systems. Incorporation of different fluorescently tagged bases by primer extension at the SNP site enables allele products to be distinguished through both their mass and emission spectra. Tests on artificial pools indicate it to be a sensitive approach. Most recently, interest in employing commercially available microarrays, originally designed for multiple SNP detection on single individuals, for analysis of pooled DNA has led to exciting preliminary conclusions. This strategy is discussed in **Subheading 3.** for that reason. We have investigated its applicability employing the Affymetrix GeneChip® Human Mapping 10K Array Xba 131, a technology designed to genotype more than 10,000 SNPs using 250 ng genomic DNA. In addition to genotyping individual DNAs, we have shown that the Affymetrix GeneChip® Human Mapping 10K Array Xba 131 can be employed accurately to estimate allele frequencies of DNA pools using the quantitative Relative Allele Signal (RAS) scores generated from the signal intensities on the microarray ***(34)***. It is highly probable that a very similar protocol will enable the interrogation of the next generation of microarrays displaying 120K SNPs spaced at roughly 20-kb intervals and hence a scan by association for the entire genome.

For this approach using the 10K GeneChip, pooled genomic DNA is first digested using *Xba*I restriction enzyme. Adaptors, containing generic sequences, are then ligated onto the digested products. Next, a single PCR primer—optimized for short fragments—attaches to the ligated adaptors and amplifies all the DNA fragments. The PCR products are then purified, fragmented, end-labelled, and hybridized to the microarray ready for scanning employing an Affymetrix SNP array (GeneChip® Mapping 10K Array Xba 131).

1.3. DNA Extraction

DNA can be prepared by standard SDS/poteinase K digestion and phenol extraction, or by application of generic spin column methods (e.g., Quiagen: QIamp DNA Blood maxi kit). All these can provide material that meets the general requirements for pooling and SNP detection. Our experience in SNP studies on pooled DNA is confined to DNA extracted from mouth swabs following exactly the procedures outlined by Freeman et al. ***(35)***. This approach provides access to material sent by post and is a cheap and efficient procedure now tested for longer than 5 yr with no contraindications.

1.4. Pool Construction

Irrespective of the type of marker to be used, the first step in pool construction is the superficially simple one of collecting and combining equal quantities of DNA from case and control samples. Depending on the method used for purification, quantification of DNA content by ultraviolet spectroscopy can lead to overestimation of the concentration and methods based on fluorimetry with a DNA-specific dye are to be preferred for their specificity. Because several dilutions may have to be undertaken, additional problems can be encountered through the pipetting of small volumes from viscous solutions. Repeated estimations of the samples are recommended. Finally, the fragment sizes of the DNA molecules are dependent on the extraction protocol, which may affect the efficiency of the PCR. It is worthwhile to check at least a subset of samples using a "real-time" PCR approach for the estimation of DNA template concentration. This measures template concentration by determining the number of PCR cycles required to reach a predetermined threshold level of product, which depends on the starting concentration of templates (*see* **ref. *5*** for further description and suggested protocol and **Note 1**).

1.5. Validation of the Pooled DNA

To ensure that the allele frequencies estimated from the pooled DNA accurately reflect the allele frequencies of the individuals comprising the pool, it is recommended that the SNP genotype profile for the DNA pool should be obtained for 10 SNPs using whichever of the SNP genotyping methodologies

Table 1
SNP Detection Technology as Applied to the Analysis of Pooled DNA Samples

SNP detection methodology	Reported sensitivity[b]	References
Electrophoresis in gels or capillaries		
Incorporation of UTP, glycosylation and alkaline cleavage (e.g., SNaPit)	Allele frequencies detected to within 1–2%	***12***
Chain termination and fluorescent tagging (e.g., SnapShotand SNuPe - see text)	Mean error of 1–1.5% in estimating differences in allele frequencies between pools[a]	***9,13***
Use of differentially fluorescent tagged primers	Allele signal ratios difficult to quantitate accurately.	***14***
	A procedure using a single tagged primer gave accuracies to within 1%	***15***
Light emission		
Bioluminometric assay coupled with modified primerextension (BAMBER)	Frequencies down to 5% level detected	***16***
Pyrophosphate coupled bioluminometric assay - Pyrosequencing	Sensitive and accurate to within 1%, (1.1 ± 0.6%)	***17–20***
Real-time (kinetic) PCR coupled to quantitation of productby SYBR Green 1 binding/ TaqMan™	Allele frequencies detected to within 1–5%	***21,22***
Microarrays		
(a) Coupling to chimeric primers combining locus specific and unique identifier tag followed by binding to oligonucleotide tag arrays.	Variable allele frequency estimates; but, generally, to within 5%	***23***

(b) Fluorescent signal quantitation on Affymetrix HuSNP arrays	Estimates of actual frequencies not attempted	***24***
High-throughput sequencing	Minor alleles detected at 5% level	***25***
Denaturing high-performance liquid chromatography (i.e., DHPLC)	Mean error of 1–1.5% in estimating differences in allele frequencies between pools[a]	***9,26,27***
Kinetic FP-TDI		
Two-color primer extension assay with real time monitoring of fluorescent polarisation	Allele frequencies to within 3.3% ± 0.85	***28***
Mass spectrosopy – (MS)		
Matrix-assisted laser desorption ionisation/time of flight (i.e., MALD-TOF) MS	Allele frequencies to within range 0.05–4%[a]	***9,29–31***
Electrospray ionization MS	Can detect alleles at 2% level	***32***
SSCP - Quantitative single-strand conformation polymorphism analysis	Allele frequencies detected with SD <1.8%	***33***

Sensitivity is dependent on allele frequency.

For most QTL surveys, alleles within the range 10 to 90% are considered to be likely targets and the sensitivities quoted are relevant to this range.

[a]Le Hellard et al. ***(9)*** compared SNaPshot, primer extension with dHPLC, and MassARRAY (MALDI-TOF) and reported standard errors of the mean to be 0.003 to 0.066, 0.003 to 0.017, and 0.003 to 0.004, respectively.

[b]The reader is referred to the original articles for information on how sensitivities/accuracies are calculated for each.

is appropriate. The allele frequencies estimated from pooled DNA can then be compared with the known SNP frequencies established by individual genotyping. For this purpose it is important that the pooled allele estimates should be corrected for preferential amplification of either allele as described (*see* **Subheading 3.1.9.** and **Note 2**).

2. Materials

2.1. SNaPshot™ to Assess SNP Frequencies in Pooled DNA

The following list of requirements and subsequent protocol is based on the analysis of a single pooled sample of DNA.

2.1.1. Amplification of Genomic DNA

1. Pooled DNA sample @10 ng/µL
2. Taq polymerase. AmpliTaq® (ABI, Applied Biosystems, Foster City, CA, cat. no. 94404) or equivalent.
3. 40 m*M* deoxynucleotides triphosphates (dATP, dCTP, dGTP, dTTP; ABgene, Epsom, UK).
4. 25 m*M* $MgCl_2$.
5. Deionized water.
6. Buffer (10X PCR reaction buffer (ABgene).
7. PCR oligonucleotide primers (10 µ*M*).
8. MJ Research thermal cycler (GRI or equivalent).
9. PCR Plates or strip tubes (ABgene).

2.1.2. PCR Purification

1. Shrimp alkaline phosphatase (SAP) 1 U/µL (USB Corporation, OH, cat. no. 44128).
2. MJ Research thermal cycler.

2.1.3. SNaPshot Reaction

1. SNaPshot multiplex ready reaction mix contained in the ABI Prism®SNaPshot multiplex kit (Applied Biosystems, Foster City, CA).
2. SNaPshot oligonucleotide primer @ 0.5 µ*M*.
3. Deionized water.
4. MJ Research thermal cycler.

2.1.4. Postextension Clean Up

1. ExoSAP-IT™ (USB corporation).
2. MJ Research thermal cycler.

2.1.5. Electrophoresis on the 3100 Genetic Analyzer

1. ABI Prism 3100 genetic Analyzer with POP-4 and 36-cm array.
2. Matrix standard Set DS-02 [dR110, dRGG, dTAMRA™, dROX™, LIZ™].

3. GeneScan™-120 LIZ™ size standard (ABI).
4. Formamide Ultra (National Diagnostics; AGTC Bioproducts Ltd, Hull, UK).
5. ABI Prism® Genotyper® 3.7 NT software (ABI).

2.2. Microarray Estimation of SNP Frequencies in Pooled DNA Samples

As the protocols employed for labeling and hybridization to the Affymetrix microarrays do not differ significantly from those outlined in the manufacturer's handbook, it would be redundant to reproduce them here. We have found, however, that there are sections of the protocol that are particularly important to follow. In addition, for one step, we have found that an alternative protocol (suggested by the manufacturer) has worked better in our hands (*see* below). A major modification in genotyping pools is the requirement to estimate allele frequency, rather than simply call homozygotes and heterozygotes; **Subheading 3.2.12.** covers this in detail.

2.2.1. Laboratory Set-Up

As noted in the manufacturer's protocol, correct laboratory set-up has been shown to be critical in the success of the Affymetrix GeneChip® Mapping Assay. It is recommended that a single direction of workflow be employed to reduce the risk of contamination, particularly with PCR products. There are three principle "areas" that should be prepared before commencing the assay.

The most important area is a "pre-PCR Clean Room" in which restriction digest and ligation stages should be conducted. This room should be free of PCR products (amplicons) and the only DNA present should be that to be used in the assay. This room also should contain any necessary reagents listed under **Subheadings 3.2.2.** and **3.2.3.** In addition to the usual precautions of wearing gowns and gloves, use of hairnets and safety masks are strongly encouraged. Individuals should not move from areas containing high amounts of PCR amplicons into the pre-PCR Clean Room directly without taking precautions to eliminate possible contaminants. A change of clothes and a shower is recommended for this purpose.

The second area is a "PCR Staging Room" in which the products from the ligation stage are prepped ready for the PCR reaction. The area should be essentially amplicon-free. It is acceptable to have the pre-PCR Clean Room and PCR-Staging room as one area, providing the workflow is unidirectional. This room (or area) should also contain any necessary reagents to set up the PCR but thermo-cycling amplification steps MUST be undertaken elsewhere (e.g., in the "Main Lab").

The "main lab" is where all subsequent reactions and steps of the procedure should take place. This area may have airborne PCR amplicons and DNA templates. To accommodate this laboratory setup, the GeneChip Mapping 10K Xba Assay kit is conveniently divided into three boxes, one for each area.

3. Methods

3.1. SNaPshot to Assess SNP Frequencies in Pooled DNA

3.1.1. Pool Construction

Genomic DNA purification and pool construction as described (**Subheadings 1.3.** and **1.4.**).

3.1.2. Primer Design

PCR oligonucleotide primers flanking the SNP region of interest can be designed using the web-based tool, Primer3 (http://frodo.wi.mit.edu/primer3/primer3_www.cgi), and should have predicted annealing temperatures (in 50 m*M* salt) of 58ºC ± 1ºC. The SNaPshot oligonucleotide primer must anneal to the complementary DNA strand directly adjacent to the SNP being interrogated. The optimal design is for a 20-mer with no extendable hairpin structures and to have annealing temperatures between 50 and 60ºC. Primers can also be designed to be complementary to the anti-sense DNA strand, if it is problematic to design a primer for the sense strand. There is a web-based tool that can be used to assist this primer design at (http://sgdp.iop.kcl.ac.uk/leo/cgi-bin/snpshot.cgi)

3.1.3. Amplification of Genomic DNA

1. Prepare the following reagents on ice to the following final concentrations with a final reaction volume of 10 µL in 1X PCR reaction buffer (add deionized water as necessary):

   2.5 m*M* $MgCl_2$
   0.8 m*M* dNTPs
   0.3 µ*M* forward PCR oligonucleotide primer
   0.3 µ*M* reverse PCR oligonucleotide primer
   1.6 U *Taq* polymerase
   Add 20 ng pooled genomic DNA (2 µL @10 ng/µL).

2. Perform amplification using a thermal cycler with the following standard conditions:

   96°C for 5 min;
   96°C for 45 s;
   62°C for 45 s; and
   72°C for 45 s;
   Decrease by 0.4 per cycle, 35 cycles
   72°C for 5 min.
   Store at 4°C if not proceeding directly to the next step.

3.1.4. PCR Purification

The PCR template has to be purified to remove unincorporated PCR components, which otherwise would interfere in the subsequent SNaPshot thermal cycling reaction.

1. Vortex the ExoSAP-IT briefly to mix.

2. Mix 5 µL of PCR product with 2 µL of ExoSAP-IT.
3. Using a thermal cycler incubate for 15 min at 37°C and follow with an enzyme inactivation step of 80°C for 15 min.

At this stage, samples can be stored at 4°C for short periods or 20°C for long-term storage.

3.1.5. SNaPshot Reaction

This follows the protocol as described in the ABI Prism SNaPshot manual, with minor modifications:

1. Combine the following reagents on ice, to a final volume of 10 µL:

   2 µL of ABI Prism SNaPshot ddNTP Ready Reaction Mix
   2 µL of purified PCR sample,
   0.25 µ*M* of SNaPshot extension primer (0.5 µL @ 0.5 µ*M*)
   5.5 µL of deionized water.
2. Mix, and then spin briefly.
3. Perform thermal cycling according to SNaPshot protocol:

   Thermal ramp (2°C/s) to 96°C;
   96°C for 10 s;
   Thermal ramp (2°C/s) to 50°C;
   50°C for 5 s;
   Thermal ramp (2°C/s) to 60°C; and
   60°C for 30 s.
   Repeat for 25 cycles.
   Maintain at 4°C if you cannot proceed directly to the next stage.

3.1.6. Postextension Clean Up

This step removes the 5' phosphoryl group of excess ddNTPs that would otherwise co-migrate with the SNP fragment of interest.

1. Add 1.0 U of SAP (1 U/µL) straight into the SNaPshot reaction mix.
2. Incubate on a thermal cycler for 37°C for 1 h followed by an enzyme inactivation step of 72°C for 15 min.

3.1.7. Electrophoresis on the 3100 Genetic Analyzer

1. Combine the following:

   0.5 µL of SAP cleaned samples.
   9.0 µL of Hi Di formamide (ref).
   0.5 µL of GeneScan-120 LIZ size standard (ABI).
2. Load samples onto ABI Prism 3100 genetic analyser with POP-4 and 36-cm array, with the following run parameters:

Dye set:	E5
Run Module:	SNP36_pop4 default
Analysis module:	GS120analysisgsp.

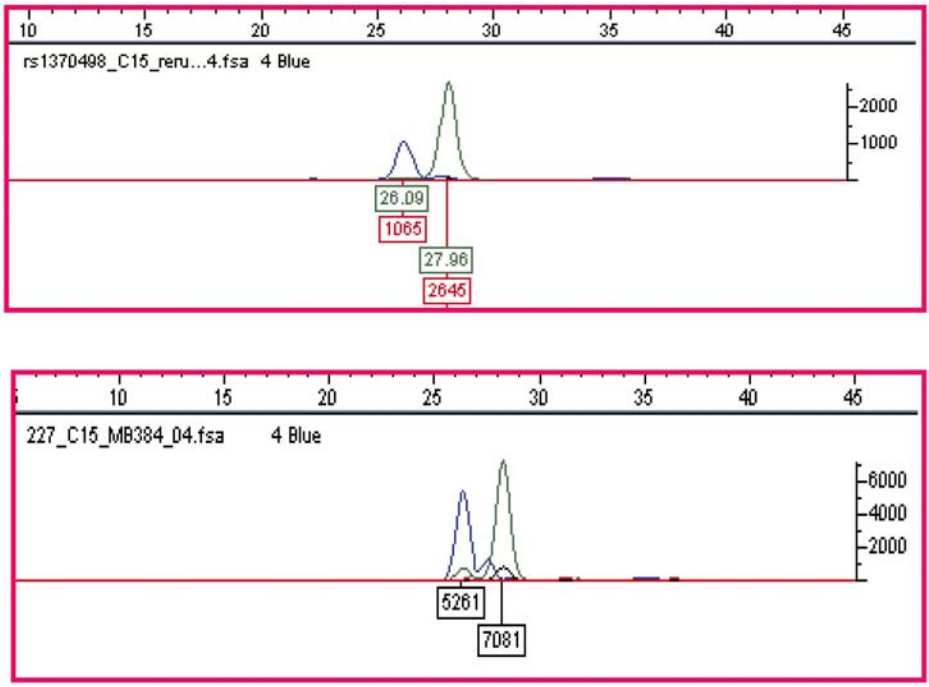


Fig. 1. SNaPshot analysis of SNP allele frequencies for case and control pooled DNA samples. Pools of 100 DNA samples from **(a)** high cognition (cases) and **(b)** mid "*g*" (controls) prepared as described in Hill et al. ***(36)***. Both panels indicate genotyping by SNaPshot for the same G/T SNP. SNaPshot SNP profile provided by Linzy Hill.

3.1.8. Data Analysis

Analyse data with GeneScan Analysis Software visualized in ABI Prism Genotyper 3.7 NT software.

3.1.9. Estimation of Allele Frequencies in the DNA Pool

The allele frequencies are reflected in the relative peak heights displayed in the electropherogram trace (*see* **Fig. 1**). If the efficiency of detection is identical for each of the individual alleles, the relative peak heights (e.g., for allele A, this is calculated as the peak height of allele A divided by the sum of the peaks heights of alleles A and B; viz. A/[A+B]) provide a direct measure of allele frequency. In practice, the efficiencies of detection for the two alleles of a SNP may differ. This could be a result of unequal amplification of heterozygotes, differential efficiencies in the incorporation of the ddNTPS and unequal emission energies for the different florescent dyes.

Examining the deviation from 50:50 for known single individual heterozygotes can assess this and a correction can be applied to estimate the true allele frequency. Given signal strengths of H_A and H_B of a pooled sample, the corrected allele frequencies (f) for alleles A and B are $f_A = H_A/(H_A + KH_B)$ and $f_B = 1 - f_A$, where K = ratio of the signal for the allele which is less well amplified (A) to that of allele B (the better amplified). *See* Sham et al. ***(5)*** for a more in-depth discussion of pool design, replication issues, application of correction functions, and other aspects of experimental design that will help to reduce technical and sampling errors. For many purposes, estimation and application of a correction coefficient, K, may be unnecessary if allele frequencies are being compared between groups, such as in case / control comparisons as it is the differences, rather than the absolute values, that are important. This particularly applies where allele frequencies are not in the extremes of their range as is likely to be the situation for significant QTLs for multi-factorial traits.

3.2. Microarray Estimation of SNP Frequencies in Pooled DNA Samples

3.2.1. Adjustment of Pooled DNA Preparations

The optimum concentration of pooled genomic DNA for use in the Affymetrix GeneChip Mapping Assay is 50 ng/µL. It is preferable that samples are diluted in reduced ethylene diamine tetraacetic acid (EDTA) TE buffer (0.1 m*M* EDTA; 10 m*M* Tris-HCL, pH 8.0) as the elevated EDTA concentration in regular TE (1.0 m*M* EDTA; 10 m*M* Tris-HCl; pH 8.0) may interfere with the assay's enzymatic reactions although use of standard TE has not proved to be a problem in our experience. A total of 250 ng DNA in a maximum volume of 15.5 µL is required for each assay; therefore, the minimum concentration of genomic DNA is approximately 16.1 ng/µL. As for individual DNA genotyping, all pooled samples must be free from contamination (with other DNA or PCR products). Although contamination of individual genomic DNAs contributing to the pool is unlikely to affect the estimation of overall allele frequency, good practice indicates that contamination of any samples should be avoided. Until complete familiarity with the system and protocols is obtained we have found it best to work with small batches—adjustment of the starting material and reagents appropriate for 8 assays has proved optimal.

3.2.2. Restriction Digest

Perform as per manufacturer's protocol, located at: https://www.affymetrix.com/support/downloads/manuals/10k_manual.pdf.

3.2.3. Ligation

Perform as per manufacturer's protocol.

3.2.4. PCR Setup

Perform as manufacturer's protocol.

3.2.5. PCR Product Purification

To assess the quality of the amplified products obtained from the digested and ligated DNA, the following steps are important:

1. Run 3 µL of each PCR product mixed with 3 L of loading dye on a 2% TBE gel at 120 V for 1 h. Three distinct bands at approx 400, 710, and 875 bp should be observed against a background of other fragments (*see* **Note 3**).
2. Proceed to **Subheading 3.2.6.**; however, if this cannot be performed directly, the sample should be stored at –20°C.

3.2.6. PCR Purification

The alternative protocol employing QIAGEN MinElute PCR Purification Kit was found to perform better as follows (*Note:* all buffers should be stored at 20–25°C):

1. Combine the four PCRs and aliquot five equal volumes to five 1.5-mL collection tubes.
2. Mix five volumes PB buffer to 1 vol PCR.
3. For each PCR, place a MinElute column in a 2-mL collection tube and stand in a suitable rack.
4. Add each PB buffer/PCR mix to center of a MinElute column and centrifuge at maximum for 1 min to bind DNA to membrane of MinElute column.
5. Discard through-flow and place each column back in the used collection tube then add 750 µL of PE buffer to each MinElute column.
6. Centrifuge at maximum for 1 min to wash DNA bound to column.
7. Discard through-flow and place the column back in its used collection tube.
8. Centrifuge at maximum for an additional minute, then discard through-flow and collection tubes. (*Note:* the additional centrifigation has been shown to be essential for recovering required yields of DNA).
9. Place MinElute Columns in clean 1.5-mL collection tubes and add 10 µL of EB Buffer to the center of each MinElute column, let each column stand for 1 min, and then centrifuge at maximum for 1 min. (*Note:* here, it is essential that EB Buffer be added to the center of the MinElute column.)
10. Collect the purified PCR products from the five tubes representing each sample and combine so each sample is now contained in a single 1.5-mL collection tube
11. Dilute 4 µL of each purified PCR sample in 156 µL of Molecular Biology Grade Water (1 in 40 dilution).
12. Quantify each purified sample applying the convention that 1 absorbance unit at 260 nm equals 50 µg/mL for dsDNA. The expected concentration should be around 11 ng/µL. At least 20 µg of purified PCR product in 45 µL is required for fragmentation. Proceed to **Subheading 3.2.7.** If fragmentation cannot be performed after PCR purification, the sample should be stored at –20°C.

3.2.7. Fragmentation

Perform as per manufacturer's protocol.

3.2.8. Labeling

Perform as per manufacturer's protocol.

3.2.9. Hybridization

Perform as per manufacturer's protocol.

3.2.10. Washing and Staining

Perform as per manufacturer's protocol.

3.2.11. Scanning

Perform as per manufacturer's protocol.

3.2.12. Estimation of Allele Frequencies (see ***Note 4***)

Allele frequency estimates are derived from RAS scores, for sense (RAS1) and antisense (RAS2) strands using GDAS software. We have routinely found that the average of RAS1 and RAS2 (RAS_{av}) provides the best estimate of allele frequency for any particular DNA pool. RAS scores should vary between 0 (for a BB homozygote) and 1.0 (for an AA homozygote), and heterozygotes should generate a relative allele signal of approx 0.5.

Depending on the efficiency of detection of the two alleles for any SNP, the RAS scores for an individual heterozygote may vary from this and a correction can be introduced to compensate for this (*see* **Subheading 3.1.9.**). Several replicates are recommended and the reader is referred to Sham et al. ***(5)*** for detailed discussion of relevant statistical points and to Butcher et al. ***(34)*** for a practical example of the application of RAS values to allele frequency estimates. Although correction of over-representative alleles is dealt with in **Subheading 3.1.9.**, correction of differential hybridization of pooled DNA specific to microarrays is dealt with in more detail in Simpson et al. ***(37)***.

4. Notes

1. Experimental errors that are unique to pooling can inflate the test statistic and so potentially contribute to an increase in false-positive associations (type 1 errors) if they are ignored. There are three main sources of variance unique to pools: 1) sampling errors (which depends on the N of the pools and the allele frequency), can be addressed by making randomly selected independent DNA pools; 2) random experimental variance in pool formation due to quantitation and pipetting errors can be addressed by making independent replicate DNA pools of the same individuals; 3) finally, measurement error in the determination of the allele fre-

quencies in the DNA pool can be addressed by repeated measurements of the same DNA pool. By incorporating such steps into the experimental design, these variances can be determined and accounted for in the test statistic.

2. To balance type I and type II errors, we recommend using a multistage replication design. For example, comparing DNA pools of extreme cases (bottom 15–25%, assuming the trait under investigation is continuously distributed) against DNA pools of extreme controls (top 15–25%). This could be followed by a within-family study that will protect against hidden population stratification, for example comparing lower versus higher DNA pools of co-twins in discordant sibling pairs. Finally, individual genotyping should be used to confirm the results.
3. It is likely that the PCR product produces a faint image. To visualize the bands it is therefore recommended that the sensitivity of the visualization software is maximized.
4. During generation of RAS scores in GDAS, the default parameters for signal detection can be lowered to increase signal detection, and thus the number of non-redundant RAS scores. As RAS scores are incorporated into an algorithm that is used to call the genotypes of individuals, it is not recommended to lower these defaults by too much. For pooled DNA however, we have found that increasing signal detection by reducing the default parameters by 25 to 50%, still yield reliable RAS scores used to estimate allele frequency. It is worth noting that RAS scores generated using default parameters are not altered when the defaults are lowered and thus remain the most accurate of the RAS scores generated.

References

1. Plomin, R., DeFries, J., Craig, I., and McGuffin, P. (2003) Behavioral genetics, *in Behavioral Genetics in the Postgenomic Era* (Plomin, R., DeFries, J., Craig, I. and McGuffin, P., eds.), APA Books, Washington, pp. 3–15.
2. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517.
3. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22,** 139–144.
4. Tabor, H. K., Risch, N. J. and Myers, R. M. (2002) Candidate gene approaches for studying complex traits: practical considerations. *Nat. Rev. Genet.* **3,** 1–7.
5. Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002) DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* **3,** 862–871.
6. Craig, I. W., and McClay, J. (2002) The role of molecular genetics in the postgenomic era, *in Behavioral Genetics in the Post-genomics Era* (Plomin, R., DeFries, J., Craig, I. and McGuffin, P., eds.), APA Books, Washington, pp. 19–40.
7. Yang, Y., Zhang, J., Hoh, J., Matsuda, F., Xu, P., Lathrop, M., and Ott, J. (2003) Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA* **100,** 7225–7230.
8. Norton, N., Williams, N. M., O'Donovan, M. C., and Owen, M. J. (2004) DNA pooling as a tool for large scale association studies in complex traits. *Ann. Med.* **36,** 146–152.

9. Le Hellard, S., Ballereau, S. J., Visscher, P. M., et al. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic. Acids. Res.* **30,** 74.
10. Breen, G., Harold, D., Ralston, S., Shaw, D. and St. Clair, D. (2000) Determining SNP allele frequencies in DNA pools. *Biotechniques* **28,** 464–470.
11. Syvanen, A. C. (2001) Accessing genetic variation; genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* **2,** 930–942.
12. Curran, S., Hill, L., O'Grady, G., et al. (2002) Validation of single nucleotide polymorphism (SNP) quantification in pooled DNA samples using SNaPIT™ technology, a glycosylase-mediated polymorphism detection method. *Biotechniques* **22,** 253–262.
13. Butcher, L. M., Meaburn, E., Dale, P. S., Schalkwyk, L., Craig, I. W., and Plomin, R. (2004) Association analysis of mild mental impairment using DNA pooling to screen 432 brain expressed single nucleotide polymorphisms. *Mol. Psychiatry* **10,** 384–392.
14. McClay, J., Sugden, K., Koch, H. G., Higuchi, S., and Craig, I.W. (2002) High-throughput single-nucleotide polymorphism genotyping by fluorescent competitive allele-specific polymerase chain reaction (SNiPTag). *Anal. Biochem.* **301,** 200–206.
15. Kirov, G., Stephens, M., Williams, N., O'Donovan, M., and Owen, M. (2000) Automated genotyping of single-nucleotide polymorphisms by extension of fluorescently labelled primers: analysis of individual and pooled DNA samples. *Balkan, J. Med. Genet.* **3,** 23–28.
16. Zhou, G-H., Kamahori, M., Okano, K., Chuan, G., Harada, K., and Kambara, H. (2001) Quantitative detection of single nucleotide polymorphisms for a pooled sample by a bioluminometric assay coupled with modified primer extension reactions (BAMPER). *Nucleic. Acids. Res.* **29,** E93.
17. Gruber, J. D., Colligan, P. B., and Wolford, J. K. (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing. *Hum. Genet.* **110,** 395–401.
18. Wasson, J., Skolnick, G., Love-Gregory, L., and Permutt, M. A. (2002) Assessing allele frequencies of single nucleotide polymorphisms in DNA pools by pyrosequencing technology. *Biotechniques* **32,** 1144–1146.
19. Nordfors, L., Jansson, M., Sandberg, G., et al. (2000) Large-scale genotyping of single nucleotide polymorphisms by Pyrosequencing™ and validation against the 5' nuclease (Taqman® assay. *Hum. Mutat.* **19,** 395–401.
20. Lavebratt, C., Sengul, S., Jansson, M., and Schalling, M. (2004) Pyrosequencing-based SNP allele frequency estimation in DNA pools. *Hum. Mutat.* **23,** 92–97.
21. Germer, S., Holland, M. J., and Higuchi, R. (2000) High-throughput SNP allele frequency determination in pooled DNA samples by kinetic PCR. *Genome. Res.* **10,** 258–266.
22. Chen, J., Germer, S., Higuchi, R., Berkowitz, G., Godbold, J., and Wetmur, J.G. (2002) Kinetic polymerase chain reaction on pooled DNA: a high-throughput, high-efficiency alternative in genetic epidemiological studies. *Cancer Epidemiol. Biomarkers. Prev.* **11,** 131–136.

23. Fan, J. B., Chen, X., Halushka, M. K., et al. (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome. Res.* **10,** 853–860.
24. Uhl, G., Liu, Q-R., Walther, W., Hess, J., and Naiman, D. (2001) Polysubstance abuse – vulnerability genes: genome scans for association, using 1,004 subjects and 1,494 single nucleotide polymorphisms. *Am. J. Hum. Genet.* **69,** 1290–1300.
25. Blazej, R. G., Paegel, B. M., and Mathies, R. A. (2003). Polymorphism ratio sequencing: a new approach for single nucleotide polymorphism discovery and genotyping. *Genome. Res.* **13,** 287–293.
26. Wolford, J.K., Blunt, D., Ballecer, C., and Prochazka, M. (2000) High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Hum. Genet.* **107,** 483–487.
27. Hoogendoorn, B., Norton, N., Kirov, G., et al. (2000b) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum. Genet.* **107,** 488–493.
28. Xiao, M., Latif, S. M., and Kwok, P. Y. (2003) Kinetic FP-TDI assay for SNP allele frequency determination. *Biotechniques.* **34,** 190–197.
29. Bansal, A., van den Boom, D., Kammerer, S., et al. (2002). Association testing by DNA poling: an effective initial screen. *Proc. Natl. Acad. Sci. USA* **99,** 16,871–16,874.
30. Werner, M., Sych, M., Herbon, N., Illig, T., König, I. R., and Wjst, M. (2002) Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectroscopy. *Hum. Mutat.* **20,** 57–64.
31. Ross, P., Hall, L. and Haff, L. A. (2000) Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. *Biotechniques.* **29,** 620–626, 628–629.
32. Zhang, S., Van Pelt, C. K., Huang, X., and Schultz, G. A. (2002) Detection of single nucleotide polymorphisms using electrospray ionization mass spectrometry: validation of a one-well assay and quantitative pooling studies. *J. Mass. Spectrom.* **37,** 1039–1050.
33. Sasaki, T., Tahira, T., Suzuki, A., et al. (2001) Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *Am. J. Hum. Genet.* **68,** 214–218.
34. Butcher, L. M., Meaburn, E., Liu, L., et al. (2004) Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Beh. Genet.* **34,** 549–555.
35. Freeman, B., Smith, N., Curtis, C., Huckett, L., Mill, J., and Craig, I. W. (2003) DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Beh. Genet.* **33,** 67–72.
36. Hill, L., Craig, I. W., Asherson, P., et al. (1999). DNA pooling and dense marker maps: a systematic search for genes for cognitive ability. *Neuroreport.* **10,** 843–848.
37. Simpson, C., Knight, J., Butcher, L., et al. (2005) Accurate allele frequency estimation from pooled DNA genotyped on microarrays. N.A.R. In press.

13

TaqMan Genotyping of Insertion/Deletion Polymorphisms

Renato Robledo, William R. Beggs, and Patrick K. Bender

Summary

The 5' fluorogenic (TaqMan) assay has been successfully used in screening for single-nucleotide polymorphisms; the very few steps required and the ability to automate each step allow for high-throughput screening. Insertion/deletion polymorphisms are an important class of markers that can be studied for different applications, such as diagnostics, genome variation, and species identification. Polymerase chain reaction (PCR) and post-PCR analysis are required to score the insertion or the deletion allele. In this chapter, we describe an expansion of the TaqMan technology for a rapid, high-throughput, screening for insertion/deletion polymorphisms in which the exact endpoints are known. The method requires minimal post-PCR analysis and can be applied to polymorphisms of any size.

Key Words: 5' fluorogenic assay; fluorogenic probes; TaqMan allelic discrimination; high-throughput genotyping; insertion/deletion polymorphisms; polymerase chain reaction.

1. Introduction

Single-nucleotide polymorphisms (SNPs) and di-, tri-, tetra-, and penta-nucleotide short tandem repeats represent the most abundant source of human genome variation. More recently, evidence exists that a different class of polymorphisms, based on the insertion/deletion (indel) of deoxyribonucleic acid (DNA) of variable length, may play a major role in establishing and maintaining human genome diversity ***(1)***. Generally, two steps are required to score the three possible genotypes (ins/ins, ins/del, del/del): a polymerase chain reaction (PCR) and a subsequent gel electrophoresis for the manual scoring of the different amplicons. The growing interest ***(2)*** in genotyping indels has motivated us to develop a simple method, based on the TaqMan technology, for fast genotyping of indel polymorphisms that does not require post-PCR analysis by gel electrophoresis nor manual scoring of the amplicons ***(3)***.

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

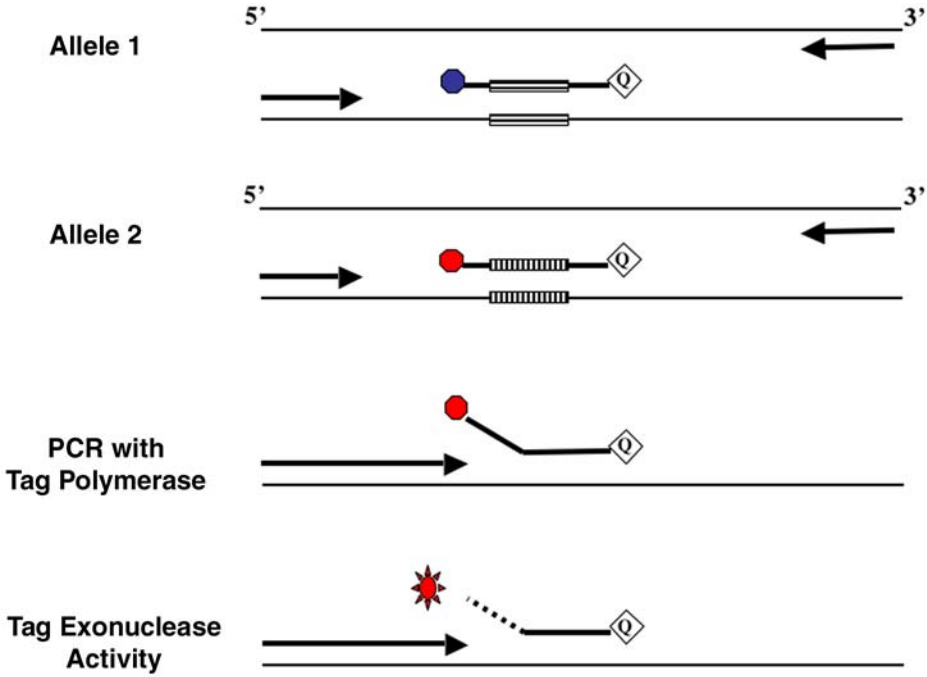


Fig. 1. Probes are indicated with quencher molecules in diamonds and reporter fluorophores as either red or blue symbols. Probe differences are indicated by striped boxes and are allele specific. The differences can be as small as a single nucleotide polymorphism. During annealing phase, probes hybridize to their specific sequences, while during the extension phase, the Taq polymerase displaces the probe and digests it releasing the reporter. In this example, samples of DNA that are homozygous for allele 1 (Al-1) would yield blue fluorescence only. DNA samples that are homozygous for allele 2 (Al-2) would yield red fluorescence only while heterozygotes would yield both blue and red fluorescence.

The TaqMan Allelic Discrimination is based on a probe technology that exploits the 5' → 3' nuclease activity of Taq DNA Polymerase ***(4)***. A TaqMan probe consists of an oligonucleotide labeled with a fluorescent reporter dye on the 5' nucleotide and a quencher dye on the 3' nucleotide (*see* **Fig. 1**). Furthermore, the 3' end of the TaqMan probe is blocked to prevent extension during PCR. When the probe is intact, proximity of the quencher to the reporter results in fluorescence resonance energy transfer suppression of the reporter fluorescence. During PCR, the TaqMan probe hybridizes to a specific target sequence in a location downstream from one of the primers. During extension, the Taq polymerase will encounter the 5' end of the probe and cleaves it during strand displacement. This action breaks the tether between the reporter and the

quencher dyes and results in increased fluorescence of the reporter (**Fig. 1**). Because this process occurs in every cycle, there is an accumulation of fluorescent reporter molecules that is a direct consequence of target amplification, and that can be measured at the end of the PCR cycles. Both primers and probe must hybridize to their targets for amplification and probe cleavage to occur. Because the fluorescence signal is generated only if the target sequence is amplified during PCR, nonspecific amplification is not detected. After the PCR, the increase of reporter fluorescence is measured in a fluorescent plate reader. If desired, the fluorescent data can be normalized to a passive reference. The passive reference is an inert fluorophore emitting at a wavelength different than that of the probe reporter. Often it is in a master mix with the probe and other components to correct for pipetting variation. It also serves to correct for variation in fluorescent detection because of well-to-well differences and differences in the optics in well-to-well measurements.

In the proposed method for genotyping indel polymorphisms, two TaqMan probes are used, each with a different fluorescent reporter. One probe recognizes a unique complementary sequence present only in the insertion, therefore identifying specifically the ins allele. The other probe is complementary to the two sequences bordering the insertion. Approximately half of the probe sequence is complementary to the 5' sequence bordering the insertion site and the remaining sequence is complementary to 3' sequences bordering the site. This probe identifies specifically the del allele, because it will hybridize to the continuous sequence that is generated by the deletion event (*see* **Fig. 2A**). After the PCR is complete allelic discrimination depends on the amount of fluorescence generated by the two probes. Fluorescence from only one probe scores the sample homozygous for whichever allele the probe targets. Fluorescence from both probes scores the sample heterozygous.

2. Materials

2.1. Reagents

1. Tris-HCl, glycerol, KCl, $MgCl_2$, and ethylenediamine tetraacetic acid were molecular biology grade.
2. 6-carboxy-X-rhodamine (ROX; (Molecular Probes, Inc.) was used as the passive reference.
3. dNTPs (Applied Biosystems).
4. AmpliTaq Gold Polymerase (Applied Biosystems).
5. Purified Genomic DNA.

2.2. Primers and Probes

1. Primers HD2a (5'-tgattctcagacagaacacac-3'), HD2b (5'-caggaggacagagttctgg-3'), and fluorogenic probe OC-2 (5'-[TET]tgatttgtcaacaacacccca[TAMRA]-3') were

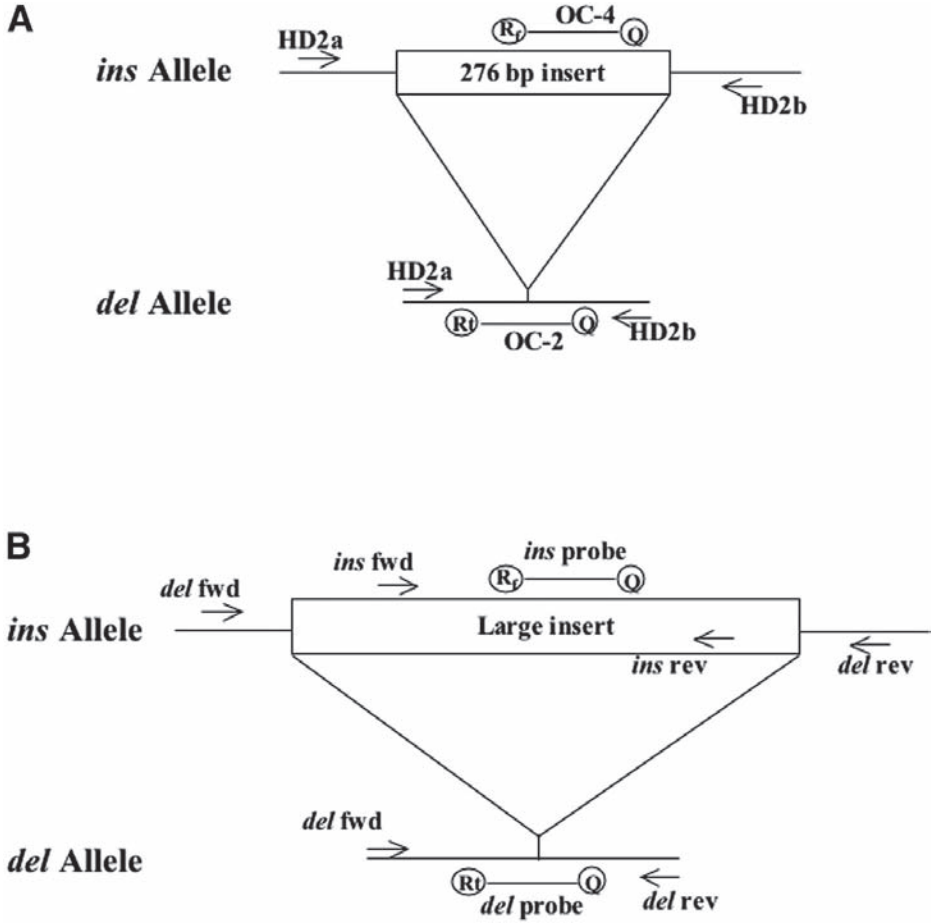


Fig. 2. Examples of probe and primer designs for genotyping indel polymorphisms of different size. **(A)** The example discussed in this article is shown with a small insert that requires only one primer pair and two probes for genotyping. **(B)** The design for larger inserts is shown. There are still two probes, one specific for insert sequences and the other specific for the immediate 3' and 5' sequences, flanking the insert. However, for larger inserts there is an additional insert-specific primer pair (ins fwd and ins rev) that amplify the portion of the insert recognized by the ins probe. In both designs the probes can be designed to hybridize to either strand, but the reporter fluorophores (R_f and R_t) must be on the 5' end with the quencher (Q) on the 3' end.

custom synthesized by Qiagen Inc. (Valencia, CA). (Note: 6-carboxy-tetrachlorofluorescein, TET; 6-carboxy-tetramethylrhodamine, TAMRA).

2. Fluorogenic probe OC-4 (5'[6FAM]ttctcgcacagatccgtcc[TAMRA]-3') was custom synthesized by Applied Biosystems (AB). (*Note:* 6-carboxy-fluorescein, 6FAM).

2.3. Equipment

1. ABI PRISM 7200 Sequence Detector (*see* **Note 1**).
2. MicroAmp Optical 96-well reaction plate and optical caps (AB). Other manufacturer's plates will work, but they should be low auto-fluorescence and the caps or sealing tape optically clear.

3. Methods

We applied the methodology to genotype a 276-bp deletion occurring on chromosome 22q13.32 ***(5)***. Depending upon the size of the indel polymorphism to be tested, either one or two primer pairs may be required, in addition to the fluorogenic probes. A single primer pair will suffice for PCR amplification of both the insertion and deletion alleles, provided that the insertion is not so large as to exceed the processivity of the Taq polymerase. For large inserts, two primer pairs are needed. One primer pair hybridizes to sequences within the insert and directs amplification of the region recognized by the ins allele probe. The other primer pair hybridizes to sequences external to and immediately flanking both sides of the indel and will identify the del allele (*see* **Fig. 2B**). In the example presented here of a 276 bp indel, only one primer pair is needed. The followings steps and guidelines should be followed for the development of the 5' nuclease assay.

3.1. Designing Probes

1. Design the probe that hybridizes to the del allele first, because the user has the least choice as to what the sequences must be. The probe has to be complementary to both the 5' and 3' sequences immediately flanking the insert. It does not have to have half of its sequences hybridizing to each side of the insert, but enough sequence from each side must be present to assure the probe will not hybridize unless the sequence is continuous as in the del allele (*see* **Note 2**). Move the sequence a few bases upstream or downstream and vary the length to obtain a probe with a melting point (T_m) between 62 and 68°C. However, the probe cannot exceed 30 nucleotides in length, and the 5' nucleotide cannot be a guanine.
2. The user has more choices for selecting the sequence of the ins allele probe because it can be any unique sequence within the insert that has a T_m within one degree of the del allele probe and fulfills the general guidelines for probe design (*see* **Note 3**).
3. In designing probes, the following guidelines should be followed: the G-C content of the probes should be in the 20 to 80% range. Avoid runs of identical nucleotides (*see* **Note 4**). Avoid a guanosine residue at the 5' end. The T_m should be 62 to 68°C (*see* **Note 5**). The probes cannot exceed 30 nucleotides in length. Select the strand that gives the probe with more cytosine than guanosine residues.

3.2. Designing Primers

1. Several restrictions should be considered in designing primers: the G-C content of each primer should be in the 20 to 80% range. Runs of identical nucleotides

should be avoided (*see* **Note 4**). The T_m of the two primers should match and be 3 to 5°C less than the probes (*see* **Note 5**).
2. If possible, within the last five nucleotides of the 3' end, there should be no more than two guanosine and/or cytosine nucleotides.
3. The forward and reverse primers cannot overlap with the probes, but can be within a few nucleotides (*see* **Note 6**).

3.3. Quantifying Primers and Probes

1. The concentration of primers and probes is determined by measuring the absorbance at 260 nm of an appropriate dilution in TE buffer. The oligonucleotide concentration (C) in μ*M* is calculated by computing the extinction coefficient factors (*see* **Note 7**), using the following formula:

Absorbance (260 nm) = sum of weighted extinction coefficients x cuvette length x oligonucleotide concentration / dilution factor

The following is an example for a probe labeled with 6-FAM and TAMRA (5'[6-FAM]tccgtcgcttgtgcagtc[TAMRA]-3') that gives an OD reading of 0.13, in a 1.0 cm cuvette, following a 1/100 dilution:

Chromophore	Molar extinction coefficient (λ_{260})	Number of chromophores	Weighted extinction coeff.
A	15,200	1	15,200
C	7050	6	42,300
G	12,010	5	60,050
T	8400	6	50,400
6-FAM	20,958	1	20,958
TAMRA	31,980	1	31,980
Sum of weighted extinction coeff.			220,888

0.13 = 220,888 M^{-1} cm^{-1} × 1.0 cm × C/100
C (conc. of probe) = 196 μ*M*.

3.4. Optimizing the PCR

1. The most critical parameters for optimal PCR results are the annealing temperature and concentrations of $MgCl_2$, primers, and probes. All these parameters have to be determined empirically. First, perform the PCR with just the primers. Fix the $MgCl_2$ at 1.5 m*M*. Vary the primers in the concentrations 0.1 μ*M*, 0.2 μ*M*, and 0.4 μ*M*, and the annealing temperature at T_m –3°C, T_m, and T_m +3°C. Fractionate the products by gel electrophoresis. If you use a template heterozygous for the indel, you may be able to score for the presence of both the del and ins fragments depending on how distal from the insert the primers positions were chosen. Inspect that the correct size fragments are present and chose the annealing temperature and primer concentration that give the best signal with the least superfluous bands.

2. Next, include one of the probes in the reaction. Titer the probe concentration at 0.05 μ*M*, 0.1 μ*M*, and 0.2 μ*M*. Use the primer concentration established in the previous experiment. Vary the annealing temperature at ±3°C from that established in the previous experiment. Vary the $MgCl_2$ concentration over the range of 3.0 m*M* to 4.5 m*M* in 0.5 m*M* increments. Include in the experiment several replicates of the template and several replicates of a no template control (water). Before the PCR, record the fluorescence of the samples at the corresponding wavelength for the probe reporter. Perform the PCR and measure the fluorescence of the samples. The no-template controls (NTCs) should have, within 10%, the same fluorescence after PCR. If the fluorescence is greater after the PCR, then suspect DNA contamination in one of the reagents. To confirm, you can electrophorese an aliquot of the NTC products on a gel. If the NTC samples are acceptable, then determine the ratio of fluorescence after PCR between the samples with template and the average of the NTC samples. The ratio should be at least twofold. Chose the conditions that give the highest ratio. If none of these conditions work, you may have to broaden the range of conditions. Try the probe at a concentration of 0.3 μ*M* and extend the annealing temperature another ±3°C.
3. Titer the second probe in the reaction at concentrations of 0.05 μ*M*, 0.1 μ*M*, and 0.2 μ*M*. Keep the other conditions and components the same as in the previous experiment using one probe. If the probes were selected with the same T_m, then the after PCR ratio of fluorescence between the template samples and the NTC will likely be greater than two for both probes. If the ratio is not greater than two, then again titer the $MgCl_2$ concentration in the reaction with both probes present.
4. The process of optimizing the reaction for primers and probes can be expedited by using a gradient thermocycler.

3.5. Prepare Master Mix

1. Once the procedure is optimized and ready to assay unknown samples, we often prepare a master mix with all of the components except the template DNA. When optimizing the conditions, a master mix can still be used; just omit from the master mix those components that are being varied and make any adjustments in volume by varying the water in the master mix. The master mix used in the example presented here, is 1.1X prepared according to the following scheme:

Reagent	Volume (μL) for a 25-μL reaction	Concentration of 1.1× mix
20% Glycerol	10	8%
10X TaqMan buffer (*see* below for preparation)	2.75	1.1X
25 m*M* $MgCl_2$	3.3	3.3 m*M*
10 m*M* dATP	0.5	200 μ*M*
10 m*M* dCTP	0.5	200 μ*M*
10 m*M* dGTP	0.5	200 μ*M*
10 m*M* dTTP	0.5	200 μ*M*

(continued)

(Table continued from previous page)

Reagent	Volume (µL) for a 25-µL reaction	Concentration of 1.1× mix
10 µ*M* each HD2a and HD2b primer mix	1.0	0.4 µ*M*
10 µ*M* OC-2, *del* allele probe (TET)	0.27	0.11 µ*M*
10 µ*M* OC-4, *ins* allele probe (6-FAM)	0.44	0.17 µ*M*
AmpliTaq Gold DNA Polymerase (5 U/µL)	0.25	0.05 U/µL
H_2O	5.00	
Total	25	

Aliquots of the Master Mix can be stored at 4°C for up to 1 wk. There are commercial sources of the Master Mix that contain all the components except the primers and probes. We have had success with Master Mixes from Applied Biosystems and Eurogentec. 10X TaqMan buffer contains: 0.1 m*M* ethylenediamine tetraacetic acid, 500 m*M* KCl, 100 m*M* Tris-HCl, and 600 n*M* ROX, pH 8.3. Mix components and filter through 0.45-µ*M* cellulose acetate filter. Stable at 4°C protected from light.

3.6. Preparing Control Reactions and Control DNA Samples

1. Prepare a 96-well plate (low auto fluorescence) with controls and sample reactions. Controls include NTC and a sample that is known to be homozygous for the del allele and a sample homozygous for the ins allele to serve as positive standards. Each 96-well plate should contain eight NTC wells, eight replicates of the ins allele standard, and eight replicates of the del allele standard, and as many as 72 unknown genomic DNA samples. **Fig. 3** shows a plate diagram with placement of control and sample reactions (*see* **Note 8**).
2. Dispense 22.5 µL of the Master Mix in each of the 96 wells in the plate.
3. Add 2.5 µL of TE buffer to wells A1-A8 for the NTC reactions.
4. Add 2.5 µL of the ins allele standard (30–60 ng) to wells A9-A12 and into wells B1–B4.
5. Add 2.5 µL of del allele standard (30–60 ng) into wells B5–B12.
6. Add 2.5 µL of each unknown sample (30–60 ng) to wells C1–H12 for the allelic discrimination assay.
7. Close the plate with Optical Caps or optically clear sealing tape.
8. Centrifuge the plate at 1000*g* for 1 min.
9. Run the PCR, according to the optimized conditions.

	1	2	3	4	5	6	7	8	9	10	11	12
A	NTC A1	NTC A2	NTC A3	NTC A4	NTC A5	NTC A6	NTC A7	NTC A8	AL1 A9	AL1 A10	AL1 A11	AL1 A12
B	AL1 B1	AL1 B2	AL1 B3	AL1 B4	AL2 B5	AL2 B6	AL2 B7	AL2 B8	AL2 B9	AL2 B10	AL2 B11	AL2 B12
C	UNKN C1	UNKN C2	UNKN C3	UNKN C4	UNKN C5	UNKN C6	UNKN C7	UNKN C8	UNKN C9	UNKN C10	UNKN C11	UNKN C12
D	UNKN D1	UNKN D2	UNKN D3	UNKN D4	UNKN D5	UNKN D6	UNKN D7	UNKN D8	UNKN D9	UNKN D10	UNKN D11	UNKN D12
E	UNKN E1	UNKN E2	UNKN E3	UNKN E4	UNKN E5	UNKN E6	UNKN E7	UNKN E8	UNKN E9	UNKN E10	UNKN E11	UNKN E12
F	UNKN F1	UNKN F2	UNKN F3	UNKN F4	UNKN F5	UNKN F6	UNKN F7	UNKN F8	UNKN F9	UNKN F10	UNKN F11	UNKN F12
G	UNKN G1	UNKN G2	UNKN G3	UNKN G4	UNKN G5	UNKN G6	UNKN G7	UNKN G8	UNKN G9	UNKN G10	UNKN G11	UNKN G12
H	UNKN H1	UNKN H2	UNKN H3	UNKN H4	UNKN H5	UNKN H6	UNKN H7	UNKN H8	UNKN H9	UNKN H10	UNKN H11	UNKN H12

Fig. 3. Example layout of a 96-well plate. Eight NTCs are in wells A1 through A8. The allele 1 standard (AL1, a known homozygote for either the ins or del allele) is in wells A9 through A12 and B1 through B4. The allele 2 standard (AL2, a known homozygote for the genotype not designated as allele 1) is in wells B5 through B12. The remaining 72 wells are the unknown samples (UNKN) to be genotyped.

10. For the example presented here the cycling conditions were as follows:

    94°C for 10 min
    Then 35 cycles of:
    94°C for 30 s,
    57°C for 1 min, and
    72°C for 1 min,
    Final extension at 72°C for 10 min.

11. For many assays, if the annealing temperature is 60°C or greater, then the annealing and extension steps can be merged into one step at the annealing temperature.

3.7. Running the Allelic Discrimination Assay

If you are analyzing the plates on an Applied Biosystems instrument, you can use the Sequence Detection Software (SDS) provided with the instrument to analyze the data. The step-by-step procedure for version 1.6 of the SDS is outlined in **Subheading 3.7.1.** If you are using a fluorometer that is not equipped with allelic discrimination software, you can analyze the data in Excel (Microsoft Corp., Seattle, WA), as described in **Subheading 3.7.2.** In either case, begin by centrifuging the plate at 1000 rpm for 1 min to remove condensation.

3.7.1. Analysis With Sequence Detection Software

1. Launch the SDS software.
2. From the File menu, choose "New Plate."
3. Choose "Allelic Discrimination."
4. Pick the appropriate fluorophore for the alleles under "Sample Type setup."
5. Define the plate wells, as shown in **Fig. 3**.
6. Click the "Show Analysis" button.
7. Click the "Post PCR Read" button. The software will perform the Plate Read.
8. From the File menu, choose "Save as" to save the plate.
9. From the Analysis menu, choose "Analyze." The computer analyzes the data.
10. From the Analysis menu, choose "Allelic Discrimination." The Allelic Discrimination window appears.
11. Check the Allelic Discrimination window and confirm that the No Amp (NTC), 1 (Allele 1 Standard), 2 (Allele 2 Standard), and 1 and 2 (heterozygous) calls have been made. **Figure 4A** shows an optimal Allelic Discrimination result.
12. An "Experimental Report" can be exported with the results and allele calls.

3.7.2. Analysis With Microsoft Excel

1. Measure the fluorescence at the two wavelengths of the ins and del allele probes.
2. Record the fluorescence and export as an Excel file.
3. Open the file in Excel, and plot the two fluorescence values as a scatter plot for all samples and controls. For example, plot the fluorescence from the ins allele probe on the *x*-axis and the fluorescence from the del allele probe on the *y*-axis (**Fig. 4B**).
4. The data will fall into four clusters: the NTC will cluster at the origin. Samples homozygous for the insert will cluster along the *x*-axis to the right of the origin. Samples homozygous for the deletion will cluster above the origin on the *y*-axis, and heterozygotes will cluster on the diagonal in the upper right quadrant.

4. Notes

1. Other instruments may be employed, like the ABI Prism 7700 Sequence Detector, or the TaqMan LS-50B PCR Detection System. Similar results should be obtained with any fluorescent plate reader that is compatible with the excitation and emission requirements of the fluorophores on the probes.
2. This is one of the possible limitations of the assay. Because the del allele probe must hybridize to the sequences immediately flanking the endpoints of the polymorphism, there is not much choice in selecting the optimum probe sequence. Moreover, it is essential that this sequence does not have homology with any other sequence located within the amplicon (**Fig. 2**). Homology of the probe with sequences located external to the amplicon will not affect the method, as explained in the following note.
3. Special care should be employed in those chromosomal regions where repeated sequences occur. In order for the methodology to be effective, it is essential that

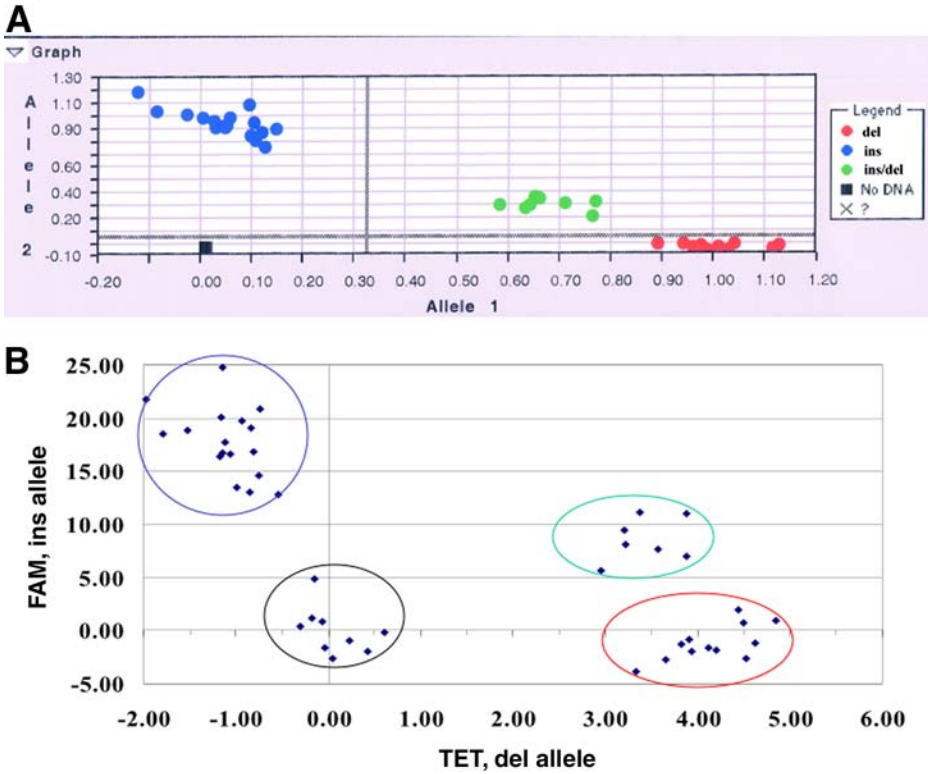


Fig. 4. Graphical outputs of genotyping results in the TaqMan assay. In **A,** the output of relative fluorescence from the SDS (version 1.6) is shown. Allele 2 was designated as the ins allele and was measured by FAM fluorescence, whereas allele 1 was designated the *del* allele and was measured by TET fluorescence. Blue dots are the samples homozygous for allele 2 and they include the allele 2 standard. Red dots are the samples homozygous for allele 1 and they include the allele 1 standard. The green dots are the samples that genotype as heterozygotes. In **B,** the graphical output using a scatter plot in Microsoft Excel is shown. The FAM and TET fluorescence data from each well were normalized by dividing by the ROX fluorescence from the same well. Fluorescent values of the NTCs were averaged for the TET and FAM channels, and these averages were subtracted from the normalized TET and FAM signals for the allele standards and the samples. The resulting fluorescent values are normalized and background corrected. These values were plotted as TET fluorescence (del allele) on the abscissa and FAM fluorescence (ins allele) on the ordinate. The data cluster into four groups. Samples circled in black around the origin are the NTCs and one sample that did not amplify. The samples circled in blue score homozygous for the ins allele. Samples circled in red score homozygous for the del allele. Sample circled in green score heterozygous.

the fluorogenic probe specific for the ins allele does not recognize a similar sequence that is located outside the insertion, but still within the amplicon (**Fig. 2**). Homology of the probe with sequences located external to the amplicon will not affect the method, because the probe will not be cleaved.

4. This is especially true for guanosine: if possible, runs of four or more guanosine nucleotides should be avoided.
5. We used Primer Express software (ABI) in calculating the T_m of primers and probes. Any program should work, just make sure to use the same program and conditions for calculating T_m values of both the primers and probes.
6. Amplicons that are less than 300 bp generally work the best. Larger amplicons can be used; however, they may require more optimization.
7. Molar extinction coefficient for other commonly used chromophores are: TET = 16,255; HEX = 31,580; VIC = 30,100; NED = 31,050; PET = 36,000.
8. Eight replicates of NTC, Allele 1, and Allele 2 standards are required to make allele calls at a 99.7% confidence level, if using the automated allele calling. Manual allele calling with less than eight replicates is possible.

Acknowledgments

We thank Drs. David Beck and Donald Coppock for helpful discussions and for revising the manuscript. **Figures 2A** and **4A** are reprinted from Robledo, R., Beggs, W., and Bender, P. (2003) A simple and cost-effective method for rapid genotyping of insertion/deletion polymorphisms, *Genomics* **82,** 580–582, with permission from Elsevier. This work was supported by funds from the NIGMS Contract N01-GM-9-2101 and the Coriell Institute for Medical Research.

References

1. Pramanik, S. and Li, H. (2002) Direct detection of insertion/deletion polymorphisms in an autosomal region by analyzing high-density markers in individual spermatozoa. *Am. J. Hum. Genet.* **71,** 1342–1352.
2. Siniscalco, M., Robledo, R., Orru, S., et al. (2000) A plea to search for deletion polymorphism through genome scans in populations. *Trends Genet.* **16,** 435–437.
3. Robledo, R., Beggs, W., and Bender, P. (2003) A simple and cost effective method for rapid genotyping of insertion/deletion polymorphisms. *Genomics* **82,** 580–582.
4. Livack, K. (1999) Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet. Anal. Biomol. Eng.* **14,** 143–149.
5. Lin, H., Pizer, E. S., and Morin, P. J. (2000) A frequent deletion polymorphism on chromosome 22q13 identified by representational difference analysis of ovarian cancer. *Genomics* **68,** 391–394.

IV

Management of Pharmacogenomic Information

14

PharmGKB

The Pharmacogenetics and Pharmacogenomics Knowledge Base

Caroline F. Thorn, Teri E. Klein, and Russ B. Altman

Summary

The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) is an interactive tool for researchers investigating how genetic variation effects drug response. The PharmGKB web site, www.pharmgkb.org, displays genotype, molecular, and clinical primary data integrated with literature, pathway representations, protocol information, and links to additional external resources. Users can search and browse the knowledge base by genes, drugs, diseases, and pathways. Registration is free to the entire research community but subject to an agreement to respect the rights and privacy of the individuals whose information is contained within the database. Registered users can access and download primary data to aid in the design of future pharmacogenetics and pharmacogenomics studies.

Key Words: Database; pharmacogenetics; pharmacogenomics; genotype; phenotype.

1. Background

In 1999, the National Institutes of Health recognized the need for a freely available collection of high-quality genotypic and phenotypic data from pharmacogenetics and pharmacogenomics studies and announced the funding of the Pharmacogenetics Research Network (PGRN). Its mission was: "to enable the formation of a series of multi-disciplinary research groups funded to conduct studies addressing research problems in pharmacogenetics. These groups are united by the purpose of developing and populating a public database, which was envisioned as a tool for all researchers in the field." ***(1)*** This tool is the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB), a knowledge base that contains genotype and phenotype data derived from individual members of the public who have agreed to allow their deidentified medical information to be used by the scientific community for pharmacogenetics and pharmacogenomics research (**Fig. 1**).

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

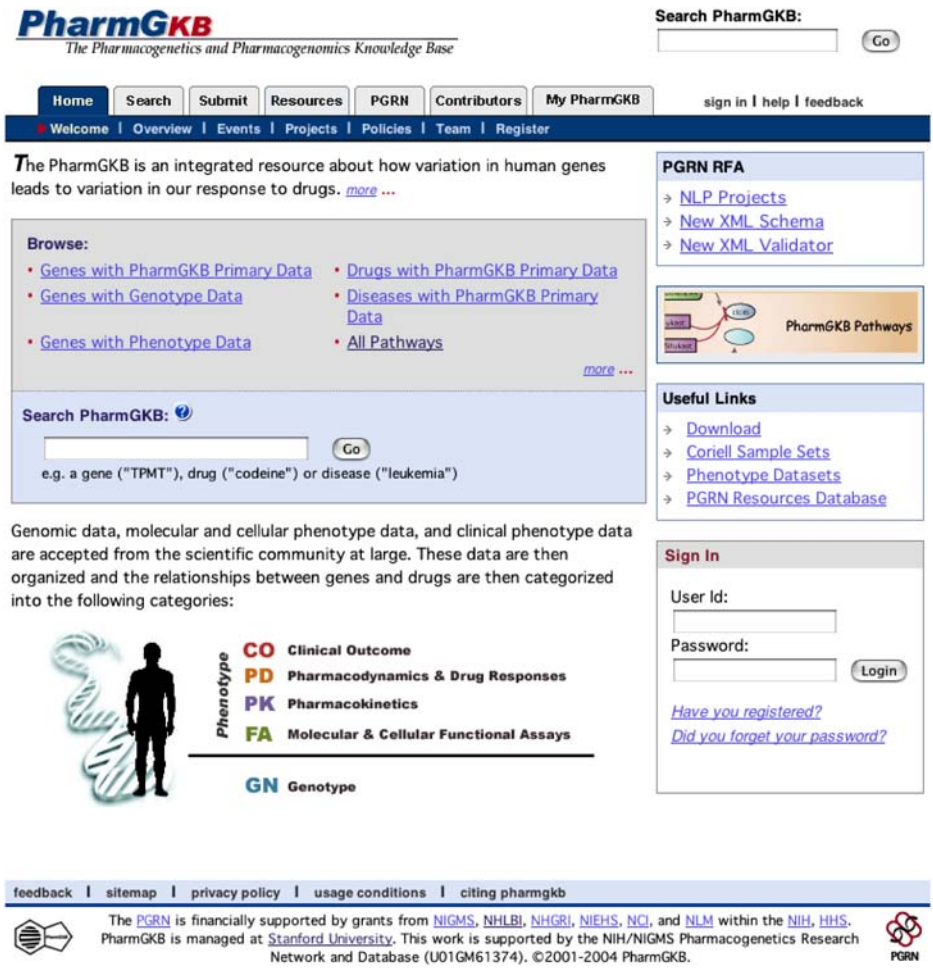


Fig. 1. The PharmGKB home page, www.pharmgkb.org, contains shortcuts to the lists of genes with genotype and phenotype data, free-text searching, or browsing by gene, drug, disease or pathway. The home page lists the categories of evidence by which data is classified in the database, Genotype, Molecular and Functional Assays, Pharmacodynamics, Pharmacokinetics and Clinical Outcomes.

2. Overview

PharmGKB stores pharmacogenetics and pharmacogenomics data in a structured format so that it can be searched, interrelated, and displayed according to the researchers interests, either for manual inspection or to download for further analyses. The knowledge base is valuable both to the researcher who is interested in a specific single-nucleotide polymorphism and its influence on a

particular drug treatment and to the researcher interested in a disease or drug and looking for candidate genes that may affect disease progression or drug response. At present, PharmGKB has data spanning a variety of disease interests and drug treatments for asthma, cancers (including breast cancer, colorectal cancer, and leukemia) cardiac arrhythmia, cardiovascular disease, and depression. The data contained within the database are from a variety of studies with different types of approaches to pharmacogenetics and pharmacogenomics, including those with a genotype to phenotype design and those with a phenotype to genotype approach, and include both high-throughput and drug pathway-focused projects ***(2)***. For example, the Pharmacogenetics of Membrane Transporters group from University of California San Francisco have used a high-throughput resequencing project to identify novel polymorphisms in transporter genes in a large group of healthy individuals of all ethnic groups and examined the effects of some of these variants on the protein functions in vitro, whereas the Pharmacogenetics of Anticancer Agents Research group have patients with differing responses to chemotherapy drugs and are identifying polymorphisms contributing to these differences.

The initial interaction with the web site is through pages devoted to genes, drugs, diseases, and pathways. These pages expand to include gene variants, methods and protocols, alleles, transcripts, proteins, and metabolites. The contents of these pages will be discussed later in this chapter. The data are represented according to a hierarchy and tagged with icons. This enables many facets of the data to be captured and stored in the database but also permits the user to find exactly what they are looking for. PharmGKB contains primary data, which are the experimentally determined data from pharmacogenetic and pharma-cogenomic studies, and Community Based Submissions of Literature Annotations and Summary data from external sources, such as the National Center for Biotechnology (NCBI) ***(3)***, Online Mendelian Inheritance in Man (OMIM) databases ***(4)***. Both primary and literature data are further classified according to Categories of Evidence and Phenotype Ontology. The categories of evidence refer to the types of measurements, both phenotypic and genotypic, conducted in that study and include genotype, molecular and functional assays, pharmacokinetics, pharmacodynamics and drug response, and clinical outcomes (*see* **Table 1** for further description).

Files can be assigned into multiple categories of evidence. To enable even more specific searches, the phenotype files also are described within the Phenotype Ontology web site (http://www.pharmgkb.org/home/project-po.jsp), which was developed by the PGRN to span the diverse types of assays or techniques and clinical tests being applied by the pharmacogenetics and pharmacogenomics research community and by MeSH terms from the National Library of Medicine Medical Subject Heading ontology ***(5)***. For example a

Table 1
Categories of Pharmacogenetic Knowledge *(18–19)*:
http://www.pharmgkb.org/resources/references/category.jsp#pd

CO Clinical Outcome

Genetic variations in the response to drugs can cause measurable differences in clinical endpoints, such as rates of cure, morbidity, side effects, and death. Data in this category demonstrate that genetic variability in the context of a drug effect significantly changes medical outcomes. These data sets are different from pharmacodynamics data sets, which may show a difference that is not sufficiently significant to alter practice or policy.

PD Pharmacodynamics and Drug Response

Genetic variation in drug targets can cause measurable differences in the response of an organism to a drug. Data in this category document that the biological or physiological response to a drug varies and that this variation can be associated with the variation of one or more genes. This variation often is measured at the whole-organism level. The measured variables may be surrogates for clinical responses, but do not constitute outcomes themselves.

PK Pharmacokinetics

Genetic variation in processes involved in the absorption, distribution, metabolism, or elimination of a drug can result in changes in drug availability. Data in this category are primarily concerned with demonstrating that genetic polymorphisms lead to variations in the levels or concentrations of drugs or their metabolites at the site of action.

FA Molecular and Cellular Functional Assays

Genetic variation can alter results of molecular and cellular funcitonal assays, and this may correlate with variations in the organism's drug response. Data in this category demonstrate associations between genetic variation and laboratory assays of function at the molecular or cellular level. These assays may test the mo lecular properties of drug targets or drug metabolizing enzymes, or may test the cellular properties of cells involved in the response to a drug (such as whole-cell gene expression).

GN Genotype

Genotype is the internally coded, heritable information carried by the organism. Variation in genotype is fundamental to pharmacogenetics and is measured as sequence variation in individual genes—the type and location of the variation, and the frequency of the variation in the populations of interest. This genetic variation is independent of individual drugs, but forms the basis for variation in response to drugs.

researcher may chose to look for any studies that observed adverse drug responses or to find all gene expression analysis experiments. The use of standardized vocabulary aids both the sorting and storage of data and supports automated methods of analysis as well as traditional human browsing. One novel aspect of the PharmGKB is that it contains linked genotype and phenotype data collected from individual human subjects. This has great potential for analysis but is also complicated by issues of ethics and security.

3. Ethics and Privacy

Although the PharmGKB is freely available to all researchers in the pharmacogenetics and pharmacogenomics community, it also has to ensure the privacy of the individuals from whom the genotypic and phenotypic data were collected. Databases that contain sensitive data, such as medical information, generally institute two types of methods of protecting privacy, scrubbing or mediation. Scrubbing, or binning, is where the data are grouped into bins containing data from several subjects with a similar range of characteristics, for example, by age range, by ethnic group, by gender, by geographic location. Mediation involves the insertion of a computer program between the user and the database that monitors the users access to the data and may limit it according to a set of rules. PharmGKB predominantly uses mediation but also uses some of aspects of scrubbing, for example, for subject ages that are binned into 10-yr ranges, little is lost in informational content but aids the deidentification process.

Summary data, such as pathways and abstracts from literature articles, always are freely accessible. However, any data relating to individual subjects are password protected. To obtain a password, it is necessary to register and accept the limitations of use. To adhere to the US government's Health Insurance Portability and Accountability Act (HIPAA) ***(6)***, PharmGKB tracks all access and use of the database. Registered users agree to this tracking and to use the data for research purposes only and not for treatment and not to attempt to identify or contact any subjects.

4. Initial Interactions With the PharmGKB Website: Gene, Drug, and Disease Pages

In PharmGKB, genes are catalogued according to the Human Genome Nomenclature Committee (HGNC *[7,8]*). In addition, alternative names and symbols also are listed and can be submitted by researchers and searched on. At present, approx 1800 genes have associated data, with approx 300 of them linked to individual subject data. The general layout of a gene page is shown in **Fig. 2**. The example shown is the *ABCB1* gene page, containing information on the gene that codes for the Multiple Drug Resistance Protein (MDR1), also known

PharmGKB
The Pharmacogenetics and Pharmacogenomics Knowledge Base

Search PharmGKB: Go

Home | Search | Submit | Resources | PGRN | Contributors | My PharmGKB | sign in | help | feedback

ABCB1
ATP-binding cassette, sub-family B (MDR/TAP), member 1

PharmGKB Accession ID: PA267
Download Data

Web Services
Comparative Genomics

Alternate Names: ATP-BINDING CASSETTE, SUBFAMILY B, MEMBER 1; ABCB1; DOXORUBICIN RESISTANCE; GP170; Homo sapiens ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), mRNA.; P glycoprotein 1/multiple drug resistance 1; P-GLYCOPROTEIN 1; PGY1; P-glycoprotein-1/multiple drug resistance-1; multidrug resistance 1

Alternate Symbols: ABC20; GP170; MDR1; NM_000927.1; P-GP; P-gp; PGY1

PharmGKB Primary Data

Variant Positions: 60 positions submitted

Phenotype Data Sets:

1	ABCB1 Functional Protein Variants				FA
	Submitted by: Kathleen M. Giacomini, PhD involving ABCB1.				
2	Irinotecan Clinical Data		PD	PK	
	Submitted by: Mark J. Ratain, MD involving ABCB1, ABCC1, ABCC2, CES1, CES2, CYP3A4, CYP3A5, TCF1, UGT1A1, UGT1A6, irinotecan, and Neoplasms.				
3	Midazolam and docetaxel clearance			PK	
	Submitted by: Mary Relling, PharmD involving ABCB1, CYP3A4, CYP3A5, docetaxel, midazolam, and Neoplastic Processes.				
4	Patient responses to tamoxifen		PD	PK	GN
	Submitted by: David Flockhart, MD, PhD involving ABCB1, CYP2C19, CYP2D6, CYP3A5, ESR1, NOS3, SULT1A1, UGT1A6, tamoxifen, and Breast Neoplasms.				
5	Pharmacokinetics of etoposide, catechol metabolite			PK	
	Submitted by: Mary Relling, PharmD involving ABCB1, CYP3A4, CYP3A5, GSTP1, UGT1A1, VDR, etoposide, and ALL (acute lymphoblastic leukemia).				
6	Pharmacokinetics of irinotecan in cancer patients			PK	
	Submitted by: Howard McLeod, PharmD involving ABCB1, ABCC1, ABCC2, ABCG2, CES1, CES2, CYP3A4, CYP3A5, UGT1A1, XRCC1, irinotecan, and Neoplasms.				
7	RNA expression in metabolism and transport genes				FA
	Submitted by: Howard McLeod, PharmD involving ABCB1, ABCC1, ABCC2, ABCG2, ADPRT, CDC45L, CES1, CES2, CYP3A4, CYP3A5, DHFR, DPYD, DPYS, DRG1, DTYMK, DUT, ECGF1, ERCC1, ERCC2, FDXR, FPGS, GGH, GSTM1, GSTP1, HMGB1, MLH1, MPO, MSH6, MTHFR, NFKB1, NME1, NME2, NT5C, POLB, POLH, PTGS1, PTGS2, RRM1, RRM2, SOD1, TDP1, TK1, TNFSF6, TOP1, TPS3, TYMS, UGT1A1, UMPK, UMPS, UNG, UP, VRS1, XPA, XRCC3, and Colonic Neoplasms.				

Legend
Φ phenotype data available
genotype data available
literature annotations available

Literature Annotations

Related Drugs	Relationship					Details
5-fluorouracil	CO				GN	
arsenite		PD			GN	
colchicine				FA		
digoxin	CO	PD	PK	FA	GN	
doxorubicin				FA		
efavirenz			PK			
Φ etoposide				FA		
Φ gemcitabine				FA		
Φ irinotecan	CO		PK	FA	GN	
Φ methotrexate	CO			FA	GN	
Φ midazolam			PK	FA		
nortriptyline		PD				
pesticides		PD			GN	
prednisone	CO	PD				
Φ rifampin				FA		
tacrolimus				FA		
vincristine				FA		
vitamin d			PK	FA		
warfarin	CO				GN	
xenobiotics			PK	FA	GN	

Related Diseases	Relationship					Details
Blood Coagulation Disorders	CO				GN	
HIV			PK			
Φ Neoplasms	CO				GN	
Parkinson Disease		PD			GN	
liver transplantation				FA		
major depression		PD				

Details
Cytogenetic Location: 7q21.1
OMIM Phenotype: Colchicine resistance
Product Name: ATP-binding cassette, sub-family B (MDR/TAP), member 1

Cross-References
LocusLink ID: 5243
OMIM Accession: 171050
UCSC Genome Browser ID: NM_000927
Ensembl ID: ABCB1
GenAtlas ID: ABCB1
GenBank ID: AF016535
GenBank ID: M14758
GeneCard ID: ABCB1
MutDB: ABCB1
Ref Seq NM Accession: NM_000927
Ref Seq NP Accession: NP_000918
Ref Seq NT Accession: NT_007933
SOURCE ID: ABCB1

feedback | sitemap | privacy policy | usage conditions | citing pharmgkb

The PGRN is financially supported by grants from NIGMS, NHLBI, NHGRI, NIDDK, NCI, and NLM within the NIH, HHS. PharmGKB is managed at Stanford University. This work is supported by the NIH/NIGMS Pharmacogenetics Research Network and Database (U01GM61374).

as P-glycoprotein. PharmGKB Primary Data links are at the top, with literature annotations below and external links in the right hand panel. The number of reported variants is highlighted, currently 60 for *ABCB1*, and links directly to the genotype data. Users can pay closer attention to genes of particular interest to them by adding the gene to their watchlist. This allows them to track when new polymorphisms are reported, both to PharmGKB and dbSNP, by checking under the "My PharmGKB" tab.

Drug and disease pages follow a similar layout style to the gene pages. Drug information including pharmacological effects, mechanisms of action and structures was obtained from VA-NDF and Apelon *(9)*. Disease information is imported from MeSH *(5)*. In the near future the scope of the database will be expanded to include pages for well known alleles, such as the Cytochrome P450 star alleles, e.g., Cyp2C9*3, and proteins, including structural representations of the locations of variants.

5. Genotype Data

The genotype data are collected and stored in the database using the eXtensible Markup Language (XML ***[10]***) and a schema developed by PharmGKB to be specific for genotyping data ***(11)***. A controlled vocabulary is used in XML to label text-based data and organize it. The PharmGKB XML schema can capture sequence information, the location of variants and exons, on the gene, alternative splicing, changes in amino acid sequence, haplotype, subject information such as gender and ethnicity, how to assay for the variant, and more than 100 other relevant data components. Recent expansion of the XML has enabled us to accept data on model organisms such as mice and rats.

Visual representation of the gene structure and location of the variants is shown using the PharmGKB "Gene Browser." The Gene Browser depicts the complete Golden Path human genomic sequence from University of California at Santa Cruz ***(12)*** in a gene-by-gene view along a continuous chromosome. The gene features such as exons, introns, promoters, etc., and their exact locations can be submitted by researchers or derived by local alignment of exonic sequences from the RefSeq database ***(13)*** to the Golden Path genomic sequence by the BLAT algorithm ***(14)***. The different gene features are displayed by color-coding as shown in **Fig. 3**. The location of reported variants are marked on the Gene Browser, by using the magnification tools the position of the variants can be depicted in a global, whole gene view or at the individual base level.

Fig. 2. *(opposite page)* The *ABCB1* gene page lists the alternate gene names, features the PharmGKB primary genotype and phenotype data in the center, literature data below, and links to external sites in the right hand panel.

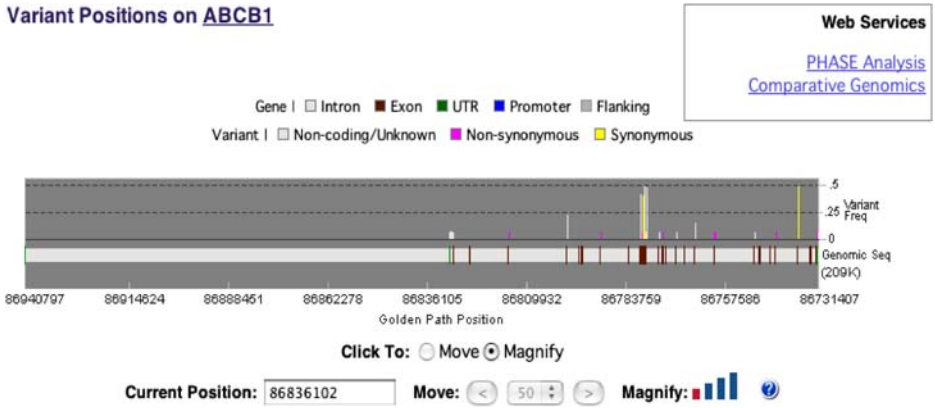


Fig. 3. The PharmGKB Gene Browser shows the relative positions of gene structures, such as exons and introns, the locations of variants, and their frequencies and properties, synonymous or nonsynonymous. The magnification tool allows the user to focus in on individual bases on the gene sequence.

The location of variants and frequencies are also represented in tables that can be downloaded in Microsoft® Excel and tab-delimited formats for use in further statistical analyses. Each variant is linked to detailed methods describing the genotyping assay performed. The frequencies of each variant can be displayed according to race or ethnicity, the user can then drill down further to obtain the list of individual subjects who were genotyped at this locus, their characteristics and genotypes at other loci. This information can also be downloaded as an Excel or tab-delimited spreadsheet or exported to run the Phase haplotype determination algorithm ***(15)***.

PharmGKB currently contains genotypes for more than 11,000 individuals at more than 5000 sites. In addition to data from PGRN research groups, we also have contributions from non-network submitters, including investigators sponsored by the National Heart, Lung and Blood Institute (NHLBI), and have specifically received supplemental funding to deposit their pharmacogenetic data in the public resource. Submitters can set the time for the release of data to the public web site to coincide with their journal publications.

6. Phenotype Data

PharmGKB accepts and displays diverse types of Phenotype Data from in vitro experiments to clinical studies. The Phenotype data files are assigned Categories of Evidence—Molecular and Functional Assays, Pharmacokinetics, Pharmacodynamics and Drug Response, and Clinical Outcomes—that describe the types of phenotypic measurements conducted in that study. In addition, the

files have been tagged with an ontology key word, designed to aid retrieval of datasets with similar types of experiments or procedures that might be of interest to the researcher, for example a researcher interested in drug-induced hypertension could find all files in which subject's blood pressures were measured.

The data within PharmGKB provides many more levels of complexity that those reported in most literature articles because for all in vivo studies, and some of the in vitro studies, the phenotype measurements are linked to individual subjects; thus, correlations of phenotypes to genotypes can be performed. Where subjects have participated in multiple studies, and genotypes have been submitted by different research groups for the same individuals, this may permit researchers to find interesting and novel genotype–phenotype relationships that had not been anticipated in the initial study hypothesis. While PharmGKB does not currently provide the analysis tools to perform these correlations, users need to be aware that it is not always wise to combine data from two different clinical studies.

Currently, 52 phenotype files are available, involving more than 3400 individuals and more than 800 types of measurements. These include both in vitro and in vivo studies, and cover all of the categories of evidence. A good representation of data on pharmacodynamics and pharmacokinetics of cancer therapies can be found, including studies on pediatric patients, as well as phenotypes associated with cardiovascular disease and hypertension. We also now accept microarray data in MAGE-ML format ***(16)***.

7. Pathways

Historically, many pharmacogenetic studies have focused on single genes involved in drug side effects, there is now a growing interest in how pathways of interacting gene can effect both drug metabolism and drug response. The CREATE group, at Washington University in St. Louis, have developed a number of drug-centered pathways to identify target genes and investigate functional polymorphisms. We have worked with them to expand some of these pathways and others to generate more interactive knowledge-linked pathways. PharmGKB pathways are drug-centered, depicting candidate genes for pharmacogenetics and pharmacogenomics studies, and they provide the means to connect separate data sets to represent the current knowledge as a cohesive snapshot. The diagrams have information content in the shape and color of the icons that represent whether the component is a gene, a drug, a metabolic intermediate, and so on. The pathways are interactive: clicking on a gene icon opens a window with the gene page, clicking on a drug opens a window of a drug page, and so on. The relationship between components of the pathway can be accessed by clicking on the yellow-headed arrows. The Irinotecan Pathway is shown in **Fig. 4** as an example.

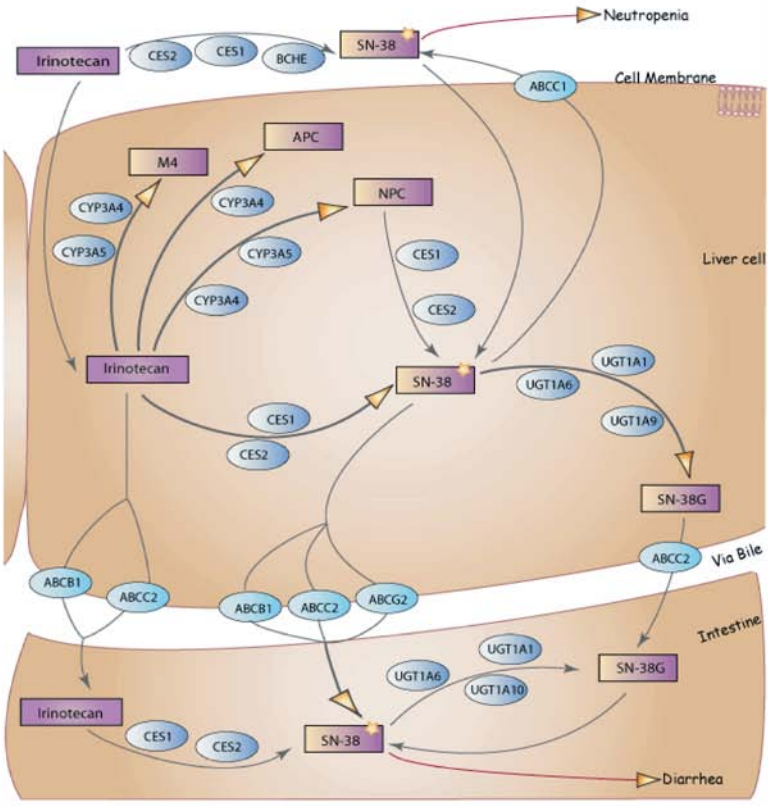
Irinotecan Pathway

Liver cell: Model human liver cell showing blood, bile and intestinal compartments, indicating tissue specific involvment of genes in the irinotecan pathway.

Legend

liver

Go



RELATED GENES
- NR1I2
- TCF1

RELATED DRUGS
- Topotecan

RELATED DISEASES
- Neoplasms
- Colonic Neoplasms
- Colorectal Neoplasms
- Gilbert Syndrome
- Carcinoma, Small Cell

DOWNLOADS
- Illustrator file (liver.ai)
- Supporting Evidence (xls)

DESCRIPTION:	This pathway shows the biotransformation of the chemotherapy prodrug irinotecan to form the active metabolite SN-38, an inhibitor of DNA topoisomerase I. SN-38 is primarily metabolized to the inactive SN-38 glucuronide by UGT1A1, the isoform catalyzing bilirubin glucuronidation. Irinotecan is used in the treatment of metastatic colorectal cancer, small cell lung cancer and several other solid tumors. There is large interpatient variability in response to irinotecan, as well as severe side effects such as diarrhea and neutropenia, which might be explained in part by genetic variation in the metabolic enzymes and transporters depicted here. Well-known variants to effect this pathway are the promoter polymorphic repeat in UGT1A1 (UGT1A1*28) and the 1236C>T polymorphism in ABCB1. While UGT1A1*28 genotype has been associated with toxicity, further evidence is needed to describe the roles of ABCB1 variants in toxicity. The effects of variants in these genes and in the carboxylesterases can be seen by clicking on the yellow-headed arrows which link to primary data from PharmGKB phenotype and genotype studies. Many of the metabolic enzymes and transporters depicted here are also involved in the pharmacokinetics of other common drugs and xenobiotics, including anticonvulsants, calcium channel blockers, macrolide antibiotics, HIV antivirals, statins and St Johns Wort, and thus co-treatment with a combination of any of these drugs may also impact efficacy and toxicity.
AUTHORS:	C.F. Thorn, M.W. Carrillo, J. Ramirez, S. Marsh, E.G. Schuetz, M.E. Dolan, F. Innocenti, M.V. Relling, H.L. McLeod and M.J. Ratain.
DATE POSTED:	September 12, 2003
DATE LAST UPDATED:	September 20, 2004

A summary is provided to describe in words the content of the graphic, its particular view and limitations, and additional, perhaps ill-defined or controversial, data that were not included in this representation. The pathways are generated by collaboration of investigators to link data, either novel or in the public domain, centered on a particular drug. The representation is a consensus of the opinions of the authors. Currently, these pathways are constructed by hand as graphic images. We are developing ways to dynamically generate pathways of equal quality from the information stored in the knowledge base.

8. Literature Annotations: The Community-Based Submissions Project

Capturing the wealth of pharmacogenetics and pharmacogenomics data already published is a considerable challenge. Most of this is stored in written natural language text in journal articles or books and not easily retrieved by automated methods. We are involved in research into natural language processing and ways in which to identify all pharmacogenetics and pharmacogenomics articles in PubMed ***(17)*** but there is still a necessity for human involvement to ensure quality data ***(18)***.

The selection of appropriate articles for populating the database ideally is performed by pharmacogenetics and pharmacogenomics researchers as they are both familiar with the fundamental literature and aware of new advances. Therefore, we have initiated a community-based submissions project, encouraging researchers to enter references to pharmacogenetics and pharmacogenomics articles, in their own specialty or otherwise, which they think would be of interest to the PharmGKB users. To date, more than 1400 literature annotations have been submitted to the database. Submitters have the ability to attach notes to an article to say why it is of particular significance and users can also leave feedback about the value or relevance of a particular article.

9. Future Directions

We aim for PharmGKB to become established as the "go-to" website for pharmacogenetics and pharmacogenomics knowledge, integrating data from current studies and literature archives, from human subjects and model organisms and providing the supporting technical and detailed protocol information. PharmGKB relies on the involvement of the pharmacogenetics and

---

Fig. 4. *(opposite page)* The Irinotecan Pathway, view of a model human liver cell showing blood, bile, and intestinal compartments, indicating tissue-specific involvement of genes in the irinotecan pathway. Drugs are depicted by purple boxes, transporter genes by turquoise ovals, and genes coding for metabolic enzymes by blue ovals. Available at: http://www.pharmgkb.org/search/pathway/irinotecan/liver.jsp

pharmacogenomics research community, both as users and submitters, and as such its continued success is dependent on cooperation and deposition of quality data. We currently provide links to resources available from PGRN network members and plan to increase our capacity for in-house analytical functionality via web services, to streamline the process for users.

Acknowledgments

The authors would like to acknowledge Dr. Michelle Whirl Carrillo, Dr. Xing Jian Lou, John Conroy, Mei Gong, Winston Gor, Tiffany Jung, Steve Lin, Feng Liu, TC Truong, Mark Woon, and Tina Zhou for their contributions to building the PharmGKB. The PharmGKB is financially supported by grants from the National Institute of General Medical Sciences (NIGMS), Human Genome Research Institute (NHGRI), National Heart, Lung and Blood Institute (NHLBI) and National Library of Medicine (NLM) within the National Institutes of Health (NIH) via the NIH/NIGMS Pharmacogenetics Research Network (UO1GM61374: Russ Altman [PI]).

References

1. NIH goals for the PGRN, http://www.nigms.nih.gov/pharmacogenetics/goals.html.
2. Altman, R. B. and Klein, T. E. (2002) Challenges for biomedical informatics and pharmacogenomics. *Annu. Rev. Pharmacol. Toxicol.* **42,** 113–133.
3. National Center for Biotechnology (NCBI); website; available at: http://www.ncbi.nlm.nih.gov/.
4. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. The Internet available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?abomim.
5. MeSH Browser, Bethesda (MD): National Library of Medicine (US), available at: http://www.nlm.nih.gov/mesh/MBrowser.html.
6. Medical Privacy - National Standards to Protect the Privacy of Personal Health Information. United States Department of Health and Human Services. Available at: http://www.hhs.gov/ocr/hipaa/.
7. Wain, H. M., Lush, M. J., Ducluzeau, F. Khodiyar, V. K., and Povey S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.* **32,** D255–2577.
8. Wain, H. M., Lovering, R. C., Bruford, E. A., Lush, M. J., Wright, M. W., and Povey S. (2002) Guidelines for Human Gene Nomenclature. *Genomics* **79,** 464–470.
9. Chute, C. G., Carter, J. S., Tuttle, M. S., Haber, M. and Brown, S. H. (2003) Integrating pharmacokinetics knowledge into a drug ontology: as an extension to support pharmacogenomics. *AMIA Annu. Symp. Proc.* 170–174.

10. Harold, E. and Means W. (2001) *XML in a Nutshell: A Desktop Reference*. O'Reilly, Cambridge, MA.
11. PharmGKB XML schema. The Internet: http://www.pharmgkb.org/schema/4.0/index.html.
12. Kent, W. J., Sugnet, C. W., Furey, T. S., et al. (2002) The human genome browser at UCSC. *Genome Res.* **12,** 996–1006.
13. The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 17, The Reference Sequence (RefSeq) Project. The Internet: http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.
14. Kent, W. J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664.
15. Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Human Genet.* **68,** 978–989.
16. Spellman, P. T., Miller, M. Stewart, J. et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3,** RESEARCH0046.
17. Rubin, D. L., Klein, T. E., and Altman, R. B. (2005) A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *J. Am. Med. Inform. Assoc.* **13,** 121–129.
18. Altman, R. B., Flockhart, D. A., Sherry, S. T., Oliver, D. E., Rubin, D. L., and Klein, T. E. (2003) Indexing pharmacogenetic knowledge on the World Wide Web. *Pharmacogenetics* **13,** 3–5.
19. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 10th ed (Hardman, J. G. and Limbird, L. E., eds.), A. Goodman Gilman, consulting ed. McGraw Hill, New York 2001, pp. 1–2.

15

Systems for the Management of Pharmacogenomic Information

Alexander Sturn, Michael Maurer, Robert Molidor, and Zlatko Trajanoski

Summary

Recent breakthroughs in biological research have been made possible by remarkable advances in high-performance computing and the establishment of a highly sophisticated information technology infrastructure. This chapter gives an overview of the main and most important technologies needed for the management of pharmacogenomic information, namely database management systems and software and hardware architectures. Because pharmacogenomics deals with a great many of public and/or proprietary data, the most prominent ways for easy storage, retrieval, analysis, and exchange are presented. Processing these data requires the use of sophisticated software architectures. Several most recent practices useful for a pharmacogenomic environment are explained. Multitiered application design and web services are discussed and described independent of the major enterprise development platforms. Because life sciences are becoming increasingly quantitative and because state-of-the-art software architectures use many system resources, this chapter presents the most recent and powerful systems for parallel data processing and data storage. Finally, shared and distributed memory systems and combinations of them as well as different storage architectures such as directly attached storage, network-attached storage, and storage-area network are explained in detail.

Key Words: High-performance computing; pharmacogenomic information infrastructure; database; database management system; software architecture; hardware architecture; web service; data warehouse; federated database system; parallel processing; data storage.

1. Introduction

There is no doubt that the sequencing and initial annotation of the human genome, completed in April 2001, is one of the great scientific advancements in history ***(1,2)***. This breakthrough in biological research was made possible by advances in high-performance computing and the use of a highly sophisticated

From: *Methods in Molecular Biology, vol. 311: Pharmacogenomics: Methods and Protocols*
Edited by: F. Innocenti © Humana Press Inc., Totowa, NJ

information technology infrastructure. High-speed computers are necessary to analyze the tens of terabytes of raw sequence data and correctly order the 3.2-billion base pairs of deoxyribonucleic acid (DNA) that compose the human genome. The assembly and initial annotation is only the first step on a long road for understanding the human genome. Many companies, research institutes, universities, and government laboratories are now rapidly moving on to the next steps: comparative genomics, functional genomics, proteomics, metabolomics, pathways, systems biology, and pharmacogenomics ***(3,4)***. The latter is the study of how an individual's genetic inheritance affects the body's response to drugs. Thus, it holds the promise that drugs might one day be tailor-made for individuals and adapted to each person's own genetic makeup. Environment, diet, age, lifestyle, and state of health all can influence a person's response to medicines, but understanding an individual's genetic makeup is thought to be the key to creating personalized drugs with greater efficacy and safety ***(5)***. Researchers are beginning the quest to determine exactly how each gene and protein functions and more important how they malfunction to trigger deadly illnesses such as heart disease, cancer, Alzheimer's and Parkinson's diseases.

Important prerequisites for pharmacogenomics or personalized medicine will be achieved by combining a person's clinical data sets with genome information-management systems. However, huge disparate data sources, like public or proprietary molecular biology databases, laboratory management systems, and clinical information management systems pose significant challenges to query and transform these data into valuable knowledge ***(6)***. The core data are collections of nucleic and amino acid sequences stored in GenBank ***(7)*** and protein structures in the Protein Data Bank ***(8)***. Additionally, this core data is used to create secondary and integrated databases, such as PROSITE ***(9)*** and InterPro ***(10)***. Furthermore, integrating data collected from high-throughput genomic technologies, like sequencing, microarrays, single-nucleotide polymorphism (SNP) detection, and proteomics, require the nontrivial development of information management systems ***(11)***. For their establishment, increasingly powerful computers and capacious data storage systems are mandatory. In the next paragraphs we will give an overview of the main and most important technologies needed for the management of pharmacogenomic information, namely database management systems (DBMS) and software and hardware architectures.

2. Databases and DBMS

Because pharmacogenomics deals with a great many of public and/or proprietary data, there is a need to easily store, retrieve, and exchange it. The major problem is the integration of the steadily increasing heterogeneous data sources.

The most prominent ways to manage and exchange bioinformatics data are as follows:

- Field/value-based flat files;
- ASN.1 (Abstract Syntax Notation One) files;
- XML files; and
- Relational databases.

Field/value-based flat files have been very commonly used in bioinformatics. Examples are the flat file libraries from GenBank, European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL), DNA Data Bank of Japan, or Universal Protein Resource (UniProt). These file types are a very limited solution because they lack referencing, vocabulary control, and constraints. In addition, on the file level, there is no inherent locking mechanism that detects when a file is being used or modified. However, these file types are primarily used for reading purposes.

ASN.1 is heavily used at the National Center for Biological Information as a format for exporting GenBank data and can be seen as a means for exchanging binary data with a description of its structure. The access concurrency is like flat files just manageable at file level, there is no support for queries, and it lacks on scalability. But because ASN.1 files convey the description of its structure, it thus provides the flexibility that the client side does not necessarily need to know the structure of the data in advance ***(12)***.

eXtensible Markup Language (XML) documents are an emerging way to interchange data and consist of elements that are textual data structured by tags. Additionally XML documents may include a Document Type Definition (i.e., DTD) that describes the structure of the elements of an XML document. XML files are hence very flexible, human readable, and provide an open framework for defining standard specifications. For example, the MGED (www.mged.org) and Gene Ontology Consortium (www.geneontology.org) have adopted XML to provide and exchange data. The weaknesses of XML are the file-based locking mechanism and the large overhead of a text-based format caused by the recurrent content-describing tags. Although XML provides query mechanisms, it lacks scalability because it does not provide scalable facilities such as indexing ***(13)***.

A relational DBMS is a collection of programs that enables to store, modify, and extract information from a relational database. Such a relational database has a much more logical structure in the way data are stored. Tables are used to represent real world objects; with each field acting like an attribute. The set of rules for constructing queries is known as a query language. Different DBMSs support different query languages, although there is a semistandardized query language called structured query language (SQL). One major advantage of the

relational model is that if a database is designed efficiently according to Codd rules ***(14)***, there should be no duplication of any data, which helps to maintain database integrity. DBMS do also provide powerful locking mechanisms to allow parallel reading and writing without data corruption.

Needless to say, there are other ways to exchange data like the Common Object Request Broker Architecture (CORBA) ***(15)***. This standard provides an intermediary object-oriented layer that handles access to the data between server and client. Another recently emerging way to exchange data is web services ***(16)***, which will be described later.

2.1. Data Warehouse and Federated Database System

Genomic management systems allow to query data assembled from different heterogeneous data sources. They are based on two different approaches:

- Data warehouse
- Federated database system

A data warehouse is a collection of data specifically structured for querying and reporting ***(17)***. Therefore, data have to be imported in regular intervals from sources of interest. These data constitute and act like a centralized repository. Applications can query these data efficaciously and create reports. Implemented data marts duplicate content in the data warehouse and allow faster responses because of a much higher granularity of the information. The drawbacks of a data warehouse are that the timeliness of the content depends on the update interval of the external data sources. These updates can be very time consuming and may result in higher storage requirements and operating costs.

Federated database systems overcome these downsides by directly accessing external data through federated database servers ***(18)***. Integration of external data can be complete (all data can be accessed) or partial (only information needed is available through the server). Shortcomings of federated databases are that queries spanning different data sources at different locations tend to be slow. Because of different query styles, dialects, and data formats federated database servers are quite complex.

The Sequence Retrieval System (SRS ***[19]***), initially developed at EMBL and the European Bioinformatics Institute, uses an interesting approach by combining the features of data warehouses and federated database systems. SRS is on the one hand heavily indexing locally stored genomic flat file databases and, on the other hand, it allows one to query DBMS on different sites. An example for a federated approach is the Mouse Federated Database of the Comparative Mouse Genomics Centers Consortium (http://www.niehs.nih.gov/cmgcc/dbmouse.htm).

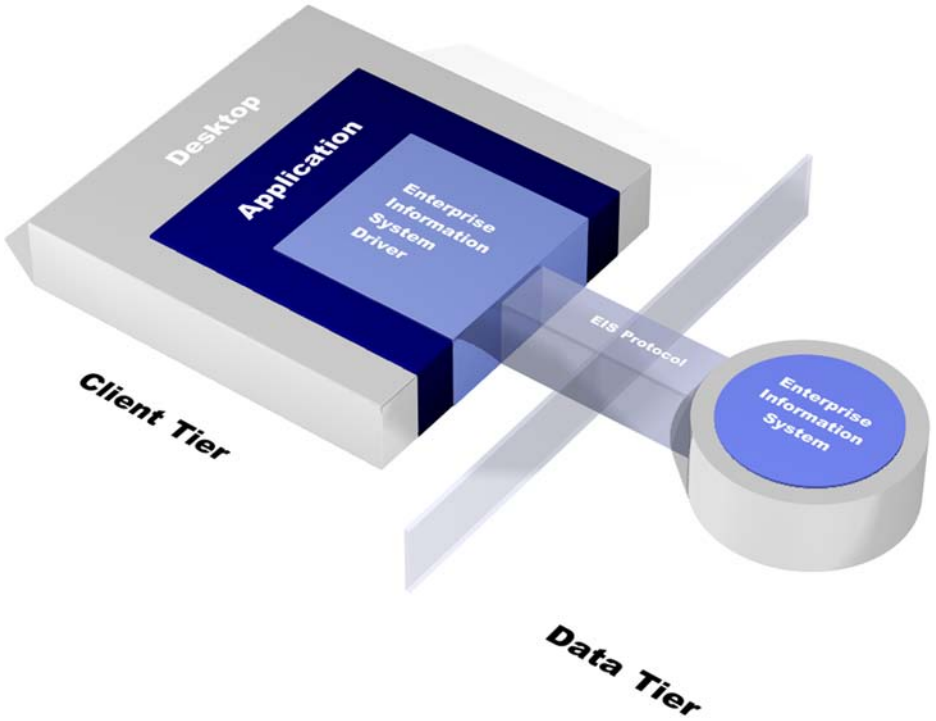


Fig. 1. Two-tier architecture. In a two-tier architecture, the application logic is implemented in the application client, which directly connects to the Enterprise Information System (Database).

3. Software Architecture

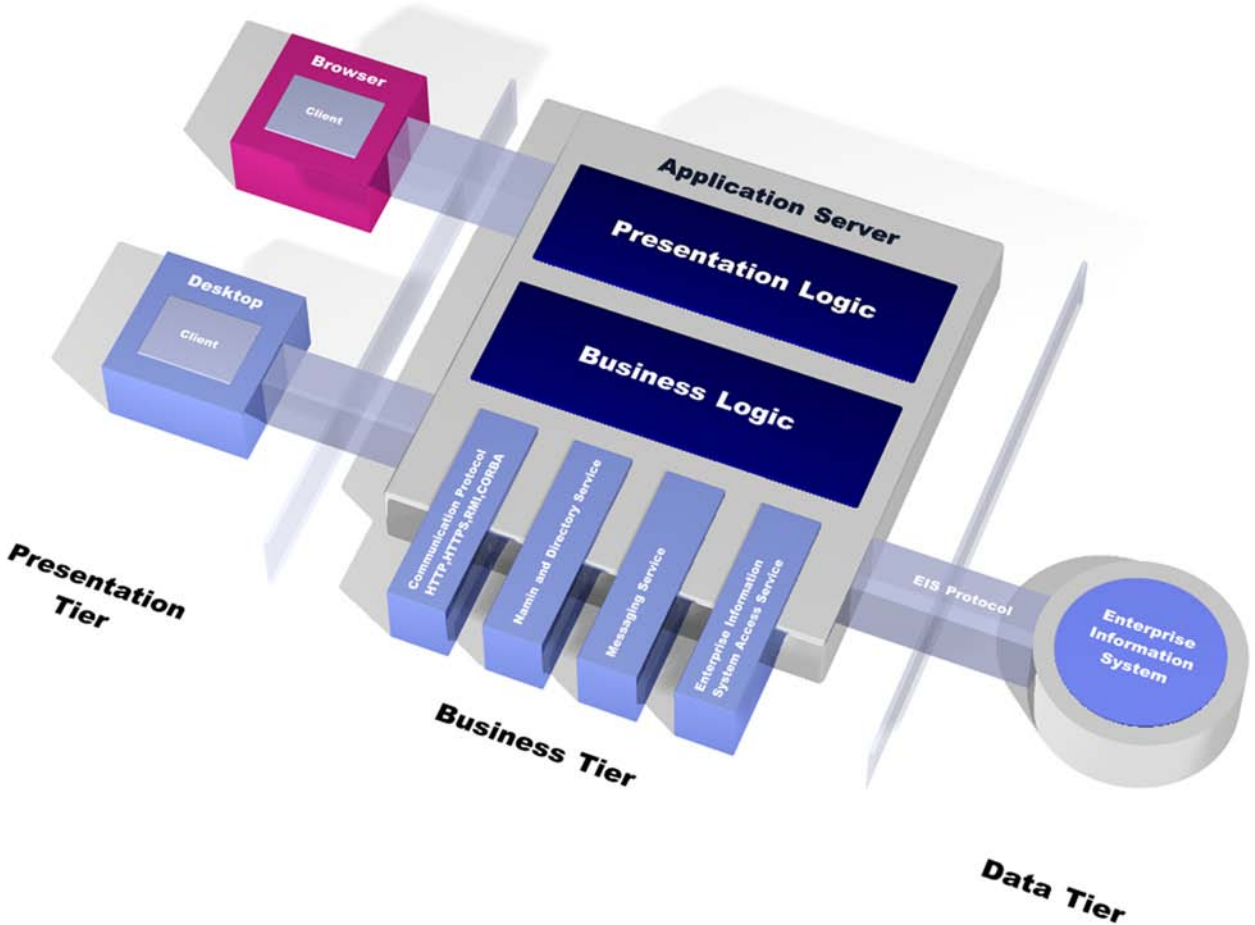
To meet the requirements of pharmacogenomic data processing systems, a sophisticated software architecture has to be used. Less complex tasks like microarray image analysis or gene expression clustering can be performed on a commonly used workstation. In this case, applications are installed locally on a client machine on which all computational tasks are performed. Required databases are either installed locally or can be accessed via the local area network (LAN) or the Internet. This kind of direct client-server access is characteristic for two-tier systems (**Fig. 1**). In a two-tier architecture the application uses the data model stored in the enterprise information system (EIS) but does not create a logical model on top of it. All the business logic is packed into the client application and, therefore, increased workstation performance is required as soon as the applications are getting more complex or computational intensive.

Furthermore, applications and database clients have to be deployed and kept up-to-date to adapt to new interfaces on the server side or to add new business logic to the system. Although there is a technology provided by Sun Microsystems® called Java Web Start to automate this cumbersome task, only a few software vendors are supporting it. In general, two-tier software application design is ideal for prototyping, for applications known to have a short lifetime, or for systems where the Application Programming Interfaces will not change. Typically, this approach is used for small applications where development costs as well as development time are intended to be low.

Most of the drawbacks of two-tier architectures can be avoided by moving to a three-tier architecture (**Fig. 2**) with an application server as central component. In a three-tier architecture, the separation of presentation, business, and data source logic becomes the principal concept ***(20)***. Presentation logic is about how to handle the interaction between the user and the software. This can be as simple as a command-line or text-based menu system, a client graphical user interface (i.e., GUI), or a HTML-based browser user interface. The primary responsibility of this layer is to display information to the user and to interpret commands from the user into actions upon the business and data source logic. The business logic contains what an application needs to do for the domain it is working with. It involves calculations based on inputs and stored data, validation of data coming from the presentation layer, and figuring out exactly what data source logic to dispatch depending on commands received from the presentation layer. The data source logic, or EIS, is about communicating with other systems that carry out tasks on behalf of the application, like transaction monitors or messaging systems. But for most applications the biggest piece of data source logic is a database, which is primarily responsible for storing persistent data. The usage of a three-tier architecture leads to the following advantages:

- Easier to modify or replace any tier without affecting the other tiers (maintenance).
- Separating the application and database functionality leads to better load balancing and therefore supports an increasing number of users or more demanding tasks.
- Adequate security policies can be enforced within the server tiers without hindering the clients.

The two major enterprise development platforms Java 2 Enterprise Edition (J2EE) and Microsoft®.Net are supporting this kind of software architecture. They can be seen as a stack of common services, like relational database access, messaging, enterprise components, or support for web services, that each platform provides to their applications. With this knowledge in the back of one's mind, the question which platform to use can be answered based on the expertise of the team members, their preferences, and based on the existing hardware and software infrastructure.



ness, and data tier. The architecture is intended to allow any of the three tiers to be upgraded or replaced independently as requirements change.

The next step in the evolution of distributed systems is web services. The concept behind is to build applications not as monolithic systems but as an aggregation of smaller systems that work together towards a common purpose. Web services are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web ***(21)***. Web services communicate using HTTP and XML and interact with any other web service using standards like Simple Object Access Protocol (SOAP), Web Service Description Language (WSDL), and Universal Description Discovery and Integration (UDDI) services, which are supported by major software suppliers. Web services are platform independent and can be produced or consumed regardless of the underlying programming language. The main limitations of web services are the network speed and round trip time latency. An additional limitation is the use of SOAP as protocol, since it is based on XML and HTTP, which degrades performance compared to other protocols like CORBA.

4. Hardware

Life science is becoming increasingly quantitative as new technologies facilitate collection and analysis of vast amounts of data ranging from complete genomic sequences of organisms to three-dimensional (3D) protein structure and complete biological pathways. As a consequence, biomathematics, biostatistics and computational science are crucial technologies for the study of complex models of biological processes. The quest for more insight into molecular processes in an organism poses significant challenges on the data analysis and storage infrastructure. Because of the vast amount of available information, data analysis on genomic or proteomic scale becomes impractical or even impossible to perform on commonly used workstations. Computer architecture, CPU performance, amount of addressable and available memory, and storage space are the limiting factors. Today, high-performance computing has become the third leg of traditional scientific research, along with theory and experimentation. Advances in pharmacogenomics are inextricably tied to advances in high-performance computing.

4.1. Parallel Processing Systems

The analysis of the humongous amount of available data requires parallel methods and architectures to solve the computational tasks of pharmacogenomic applications in reasonable time ***(22)***. State of the art technology comprises three different approaches to parallel computing:

- Shared memory systems
- Distributed memory systems
- Combination of both systems

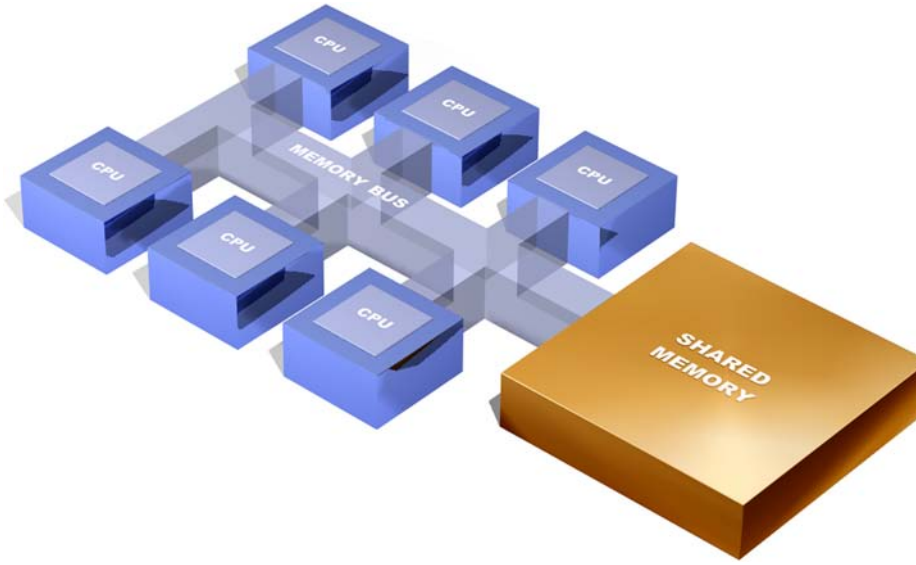


Fig. 3. Shared memory systems. A shared memory system consists of multiple processors that are able to access a large central memory directly through a very fast bus system.

4.1.1. Shared Memory Systems

In shared memory systems multiple processors are able to access a large central memory (e.g., 16, 32, 64 GBytes) directly through a very fast bus system (**Fig. 3**). This architecture enables all processors to solve numerical problems sharing the same dataset at the same time. The communication between processors is performed using the shared memory pool with efficient synchronization mechanisms making theses systems very suitable for programs with rich inter-process communication. Limiting factors are the relative low number of processors that can be combined and the high costs.

4.1.2. Distributed Memory Systems

In general, these systems consist of clusters of computers, so called nodes, which are connected via a high-performance communication network (**Fig. 4**). Using commodity state-of-the-art calculation nodes and network technology, these systems provide a very cost efficient alternative to shared memory systems for dividable, numerical computational intensive problems that have a low communication/calculation ratio. On the contrary, problems with high inter-processor communication demands can lead to network congestion, which is decreasing the overall system performance. If more performance is

Fig. 4. Distributed memory systems. In a distributed memory architecture, the various computing devices (e.g., PCs) have their own local memory and perform calculations on distributed problems. Input data and results are exchanged via a high-performance interprocess communication network.

needed, this architecture can easily be extended by attaching additional nodes to the communication network.

4.1.3. Grid Computing

Grid computing is an emerging technology, poised to help the life science community manage their growing need for computational resources. A compute grid is established by combining diverse heterogeneous high-performance computing systems, specialized peripheral hardware, PCs, storage, applications, services, and other resources placed over various locations into a virtual computing environment. For every numerical problem the appropriate computing facility in a worldwide resource pool can be harnessed to contribute to its solution. A computing grid differs from the earlier described cluster topology mainly by the fact that there is no central resource management system. In a grid every node can have its own resource management system and distribution policy. Grid technologies promise to change the way complex life science

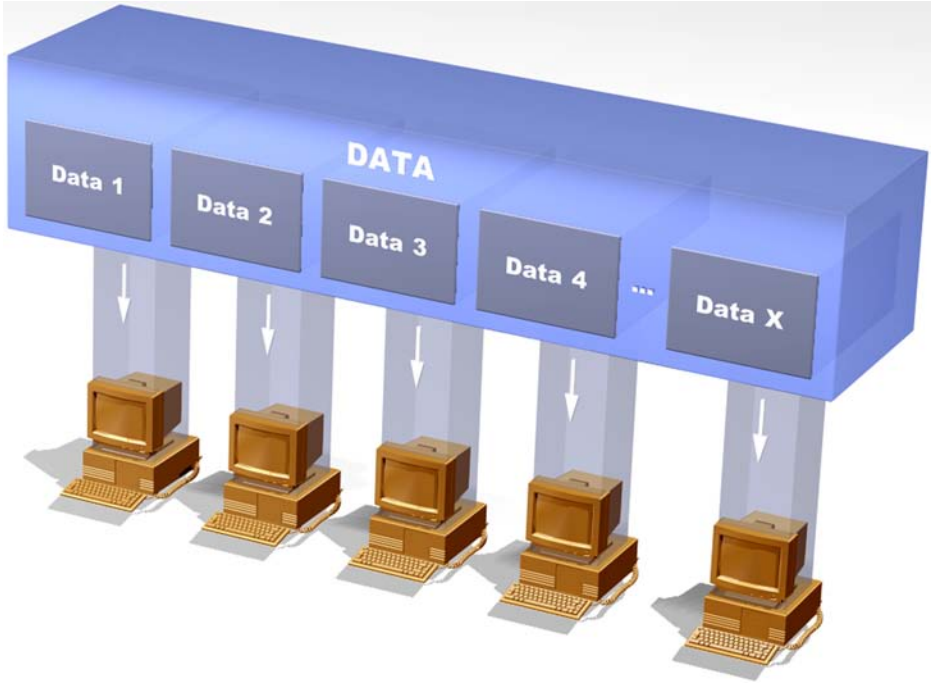


Fig. 5. Domain decomposition. Domain or data decomposition is a computational paradigm in which data to process are distributed and processed on different nodes.

problems are tackled and help to make better use of existing computational resources ***(23)***. Soon, a life scientist will look at the grid and see essentially one large virtual computer resource built upon open protocols with everything shared: applications, data, processing power, storage, and so on, all through a network.

4.1.4. Partitioning

To use the parallel features of a high performance computing facility, the software has to meet parallel demands, too. A numerical problem that has to be solved in parallel must be divided into subproblems that can be subsequently delegated to different processors. This partitioning procedure can be done either with so-called *domain decomposition* (**Fig. 5**) or *functional decomposition* (**Fig. 6**).

The term domain decomposition describes the approach to partition the input data and to process the same calculation on each available processor. Most of the parallel-implemented algorithms are based on this approach dividing the genomic databases into pieces and calculating, for instance, the sequence

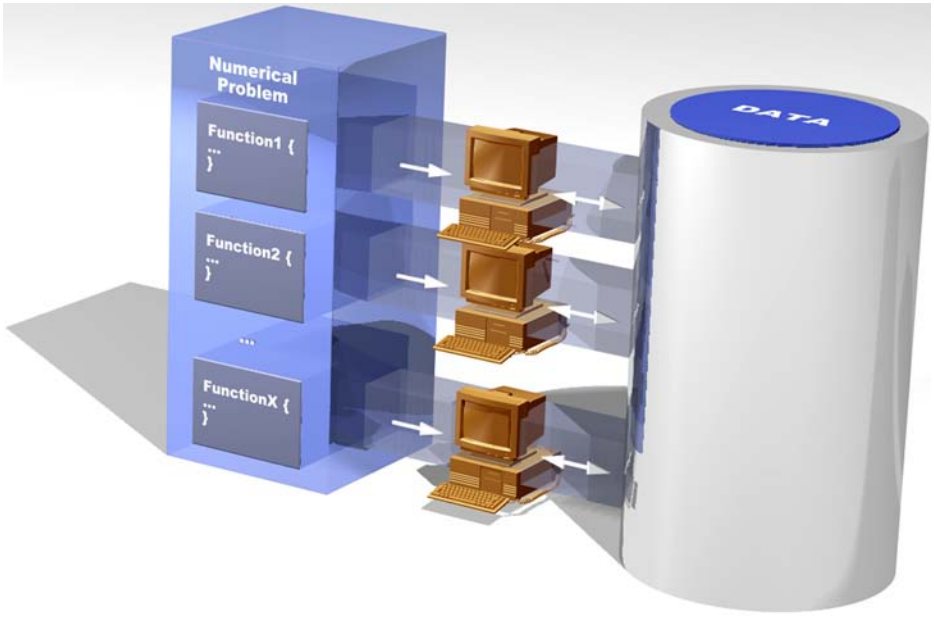


Fig. 6. Functional decomposition. Functional decomposition divides the computational problem in functional units, which are distributed onto different working nodes processing the same data.

alignment of a given sequence on a subpart of the database. The second and simplest way to implement the domain decomposition on a parallel computing system is to take sequentially programmed applications and execute them on different nodes with different parameters. An example is to run the well-known basic local alignment search tool (BLAST *[24]*) with different sequences against one database by giving every node another sequence to calculate. This form of application parallelization is called swarming and does not need any adaptation of existing programs.

On the other hand, functional decomposition is based on the decomposition of the computation process. This can be done by discovering disjoint functional units in a program or algorithm and sending these subtasks to different processors (**Fig. 6**). Finally, in some parallel implementations combinations of both techniques are used, so that functional-decomposed units are calculating domain-parallelized sub-tasks.

4.2. Data Storage

Drug discovery-related data storage and information management requirements are doubling in size every 6 to 8 mo, more than twice as fast as Moore's

Law predictions for microprocessor transistor counts. For life science organizations, data is necessary, but not sufficient for organizational success. They must generate information—meaningful, actionable, organized, and reusable data. Data must be stored, protected, secured, organized, distributed, and audited, all without interruption.

State-of-the-art storage architecture comprises the following solutions:

- Directly-attached storage (DAS)
- Network-attached storage (NAS)
- Storage-area networks (SAN)
- Internet SCSI (iSCSI)

4.2.1. Directly Attached Storage

This historically first and very straightforward method can be seen today in every PC: hard disks, floppy disks, CD-ROMs or DVDs are attached directly to the main host using short internal cables. Although in the mainframe arena storage devices, hard disks or tape drives are separate boxes connected to a host, this configuration is from a functional perspective equivalent to standard PC technology. DAS is optimized for single, isolated processor systems and small data volumes delivering good performance at low initial costs.

4.2.2. Network-Attached Storage

Network-attached storage (NAS) is defined as storage elements that are connected to a network providing file access services to computer systems. These devices are attached directly to the existing LAN using standard TCP/IP protocols. NAS systems have intelligent controllers built in, which are actually small servers with stripped operating systems, to exploit LAN topology and grant access to any user running any operating system. Integrated NAS appliances are discrete pooled disk storage subsystems, optimized for ease-of-management and file sharing, using lower-cost, IP-based networks.

4.2.3. Storage-Area Networks

A storage-area network (SAN) is defined as a specialized, dedicated high-speed network whose primary purpose is the transfer of data between and among computer systems and storage elements. Fibre Channel is the de facto SAN standard network protocol, although other network standards, like iSCSI, could be used. SAN is a robust storage infrastructure, optimized for high performance and enterprise-wide scalability.

4.2.4. Internet SCSI (iSCSI)

SCSI is a collection of standards that define I/O buses primarily intended for connecting storage subsystems or devices to hosts through host bus adapters.

iSCSI is an new emerging technology and is based on the idea of the encapsulation of SCSI commands in TCP/IP (most widely used protocol to establish a connection between hosts and exchange data) packages and sending them through standard IP based networks. With this approach iSCSI storage elements can exist anywhere on the LAN and any server talking the iSCSI protocol can access them.

5. Conclusions

A pharmacogenomic DBMS has to combine public and proprietary genomic databases, clinical data sets, and results from high-throughput screening technologies. Currently, the most important public available biological databases require disk space in the magnitude of 1 terabyte (1000 gigabytes). Considering the exponential growth of data, it can be expected that the storage requirements for proteomics will claim petabytes (1000 terabytes). Even more, systems for personalized medicine will be in the range of exabytes (1000 petabytes). Assuming that the storage capacity doubles every year it is imaginable that in 10 yr working with petabytes will be a standard procedure in many institutions. To facilitate the management, handling, and processing of this vast amount of data, such systems should comprise data-mining tools embedded in a high-performance computing environment using parallel processing systems, sophisticated storage technologies, network technologies, database and DBMS, and application services. Integration of patient information management systems with genomic databases as well as other laboratory and patient-relevant data will represent significant challenges for designers and administrators of pharmacogenomic information management systems. Unfortunately, the lack of international as well as national standards in clinical information systems will require the development of regional specific systems. Additionally all arising security issues concerning the sensitivity of certain types of information have to be solved in a proper manner. To accomplish all this stated issues, considerable endeavors have to be undertaken to provide the necessary powerful infrastructure to fully exploit the promises of the postgenomic era.

Acknowledgments

The authors express their appreciation to the staff of the Institute for Genomics and Bioinformatics for valuable comments and contributions. This work was supported by bm:bwk, GEN-AU:BIN, Bioinformatics Integration Network.

References

1 Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

2. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
3. Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003) A vision for the future of genomics research. *Nature* **422,** 835–847.
4. Forster, J., Gombert, A. K., and Nielsen J. (2002) A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol. Bioeng*. **79,** 703–712.
5. Mancinelli, L., Cronin, M., and Sadee W. (2000) Pharmacogenomics: the promise of personalized medicine. *AAPS PharmSci*. **2,** E4–E4.
6. Boguski, M. S. and McIntosh, M. W. (2003) Biomedical informatics for proteomics. *Nature* **422,** 233–237.
7. Benson, D. A., Boguski, M. S., Lipman, D. J., and Ostell, J (1997). GenBank. *Nucleic Acids Res*. **25,** 1–6.
8. Kanehisa, M. and Bork P. (2003) Bioinformatics in the post-sequence era. *Nat. Genet*. **33(Suppl),** 305–310.
9. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch A. (2002) The PROSITE database: its status in 2002. *Nucleic Acids Res*. **30,** 235–238.
10. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2003) The InterPro Database 2003 brings increased coverage and new features. *Nucleic Acids Res*. **31,** 315–318.
11. Stein, L. (2002) Creating a bioinformatics nation. *Nature* **417,** 119–120.
12. Steedman, D. (1993) *ASN 1 The Tutorial and Reference*. Technology Appraisals, Twickenham, UK.
13. Achard, F., Vaysseix, G., and Barillot, E. (2001) XML, bioinformatics and data integration. *Bioinformatics*. **17,** 115–125.
14. Codd, E. M. (1990) *The Relational Model for Data Base Management: Version 2*, Addison Wesley,
15. Hu, J., Mungall, C., Nicholson, D., and Archibald A. (1998) Design and implementation of a CORBA-based genome mapping system prototype. *Bioinformatics* **14,** 112–120.
16. Stein, L. D. (2003) Integrating biological databases. *Nat. Rev. Genet*. **4,** 337–345.
17. Kimball, R. (1996) *The Data Warehouse Toolkit: Practical Techniques For Building Dimensional Data Warehouses*. John Wiley and Sons, New York.
18. Sheth, A. P. and Larson, J. A. (1990) Federated Database Systems for managing distributed, heterogenous and autonomous databases. *ACM Computing Surv*. **22,** 183–236.
19. Zdobnov, E. M., Lopez, R., Apweiler, R., and Etzold, T. (2002) The EBI SRS server-recent developments. *Bioinformatics* **18,** 368–373.
20. Fowler, M., et al. (2002) Patterns of Enterprise Application Architecture. Addison Wesley.
21. Thallinger, G. G., Trajanoski, S., Stocker, G., and Trajanoski, Z. (2002) Information management systems for pharmacogenomics. *Pharmacogenomics* **3,** 651–667.
22. Buyya, R. (1999) *High Performance Cluster Computing: Architectures and Systems* (Vol. 1 and 2), Prentice Hall, NJ.

23. Avery, P. (2002) Data Grids: a new computational infrastructure for data-intensive science. *Philos. Transact. Ser A. Math Phys. Eng. Sci.* **360,** 1191–1209.
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Index