METHODS IN MOLECULAR MEDICINE™

Clinical Bioinformatics

Edited by

Ronald J. A. Trent

HUMANA PRESS

Clinical Bioinformatics

METHODS IN MOLECULAR MEDICINE™

John M. Walker, SERIES EDITOR

142. **New Antibiotic Targets,** edited by *W. Scott Champney, 2008*
141. **Clinical Bioinformatics,** edited by *Ronald J. A. Trent, 2008*
140. **Tissue Engineering,** edited by *Hansjörg Hauser and Martin Fussenegger, 2007*
139. **Allergy Methods and Protocols,** edited by *Meinir G. Jones and Penny Lympany, 2007*
138. **Neutrophil Methods and Protocols**, edited by *Mark T. Quinn, Frank R. DeLeo, and Gary M. Bokoch, 2007*
137. **Microtubule Protocols**, edited by Jun Zhou, *2007*
136. **Arthritis Research:** *Methods and Protocols, Vol. 2,* edited by *Andrew P. Cope, 2007*
135. **Arthritis Research:** *Methods and Protocols, Vol. 1,* edited by *Andrew P. Cope, 2007*
134. **Bone Marrow and Stem Cell Transplantation**, edited by *Meral Beksac, 2007*
133. **Cancer Radiotherapy**, edited by *Robert A. Huddart and Vedang Murthy, 2007*
132. **Single Cell Diagnostics:** *Methods and Protocols,* edited by *Alan Thornhill, 2007*
131. **Adenovirus Methods and Protocols, Second Edition, Vol. 2:** *Ad Proteins, RNA, Lifecycle, Host Interactions, and Phylogenetics,* edited by *William S. M. Wold and Ann E. Tollefson, 2007*
130. **Adenovirus Methods and Protocols, Second Edition, Vol. 1:** *Adenoviruses, Ad Vectors, Quantitation, and Animal Models,* edited by *William S. M. Wold and Ann E. Tollefson, 2007*
129. **Cardiovascular Disease:** *Methods and Protocols, Volume 2: Molecular Medicine,* edited by *Qing K.Wang, 2006*
128. **Cardiovascular Disease:** *Methods and Protocols, Volume 1: Genetics,* edited by *Qing K. Wang, 2006*
127. **DNA Vaccines:** *Methods and Protocols, Second Edition,* edited by *Mark W. Saltzman, Hong Shen, and Janet L. Brandsma, 2006*
126. **Congenital Heart Disease:** *Molecular Diagnostics,* edited by *Mary Kearns-Jonker, 2006*
125. **Myeloid Leukemia:** *Methods and Protocols,* edited by *Harry Iland, Mark Hertzberg, and Paula Marlton, 2006*
124. **Magnetic Resonance Imaging:** *Methods and Biologic Applications,* edited by *Pottumarthi V. Prasad, 2006*
123. **Marijuana and Cannabinoid Research:** *Methods and Protocols,* edited by *Emmanuel S. Onaivi, 2006*
122. **Placenta Research Methods and Protocols:** *Volume 2,* edited by *Michael J. Soares and Joan S. Hunt, 2006*
121. **Placenta Research Methods and Protocols:** *Volume 1,* edited by *Michael J. Soares and Joan S. Hunt, 2006*
120. **Breast Cancer Research Protocols**, edited by *Susan A. Brooks and Adrian Harris, 2006*
119. **Human Papillomaviruses:** *Methods and Protocols,* edited by *Clare Davy and John Doorbar, 2005*
118. **Antifungal Agents:** *Methods and Protocols,* edited by *Erika J. Ernst and P. David Rogers, 2005*
117. **Fibrosis Research:** *Methods and Protocols,* edited by *John Varga, David A. Brenner, and Sem H. Phan, 2005*
116. **Inteferon Methods and Protocols**, edited by *Daniel J. J. Carr, 2005*
115. **Lymphoma:** *Methods and Protocols,* edited by *Timothy Illidge and Peter W. M. Johnson, 2005*
114. **Microarrays in Clinical Diagnostics**, edited by *Thomas O. Joos and Paolo Fortina, 2005*
113. **Multiple Myeloma:** *Methods and Protocols,* edited by *Ross D. Brown and P. Joy Ho, 2005*
112. **Molecular Cardiology:** *Methods and Protocols,* edited by *Zhongjie Sun, 2005*
111. **Chemosensitivity:** *Volume 2, In Vivo Models, Imaging, and Molecular Regulators,* edited by *Rosalyn D. Blumethal, 2005*
110. **Chemosensitivity:** *Volume 1, In Vitro Assays,* edited by *Rosalyn D. Blumethal, 2005*
109. **Adoptive Immunotherapy:** *Methods and Protocols,* edited by *Burkhard Ludewig and Matthias W. Hoffman, 2005*
108. **Hypertension:** *Methods and Protocols,* edited by *Jérôme P. Fennell and Andrew H. Baker, 2005*
107. **Human Cell Culture Protocols,** *Second Edition,* edited by *Joanna Picot, 2005*
106. **Antisense Therapeutics,** *Second Edition,* edited by *M. Ian Phillips, 2005*
105. **Developmental Hematopoiesis:** *Methods and Protocols,* edited by *Margaret H. Baron, 2005*
104. **Stroke Genomics:** *Methods and Reviews,* edited by *Simon J. Read and David Virley, 2004*
103. **Pancreatic Cancer:** *Methods and Protocols,* edited by *Gloria H. Su, 2004*
102. **Autoimmunity:** *Methods and Protocols,* edited by *Andras Perl, 2004*
101. **Cartilage and Osteoarthritis:** *Volume 2, Structure and In Vivo Analysis,* edited by *Frédéric De Ceuninck, Massimo Sabatini, and Philippe Pastoureau, 2004*
100. **Cartilage and Osteoarthritis:** *Volume 1, Cellular and Molecular Tools,* edited by *Massimo Sabatini, Philippe Pastoureau, and Frédéric De Ceuninck, 2004*
99. **Pain Research:** *Methods and Protocols,* edited by *David Z. Luo, 2004*
98. **Tumor Necrosis Factor:** *Methods and Protocols,* edited by *Angelo Corti and Pietro Ghezzi, 2004*

METHODS IN MOLECULAR MEDICINE™

Clinical Bioinformatics

Edited by

Ronald J. A. Trent

Department of Molecular and Clinical Genetics
University of Sydney Central Clinical School
Royal Prince Alfred Hospital, Camperdown, New South Wales,
Australia

HUMANA PRESS TOTOWA, NEW JERSEY

Editor
Ronald J. A. Trent
Department of Molecular and Clinical Genetics
University of Sydney Central Clinical School
Royal Prince Alfred Hospital
Camperdown, New South Wales, Australia

Series Editor
John M. Walker
University of Hertfordshire
Hatfield, Herts., UK

ISBN: 978-1-58829-791-4 e-ISBN: 978-1-60327-148-6

Library of Congress Control Number: 2007933471

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Cover illustration: Figure 4, Chapter 16, “Online Resources for the Molecular Contextualization of Disease,” by Chi N. I. Pang and Marc R. Wilkins. Image source: www.proteinatlas.org.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

A challenge in clinical medicine is dealing with an ever-increasing volume of information. This is particularly so in the emerging "omics" era, and impacts researchers, health professionals, and the broader community. To respond to this challenge requires computer-based storage, processing, and dissemination (i.e., bioinformatics). In this volume of *Methods in Molecular Medicine*™, a number of strategies utilizing clinical bioinformatics are described. This series of articles focuses on software applications that can be used to translate information into outcomes of clinical relevance. The six themes covered include:

Gene discovery—Chapters 1 to 4.

Gene function (microarrays)—Chapters 5 to 9.

DNA mutation analysis—Chapters 10 to 12.

Proteomics—Chapters 13 to 15.

Online approaches and resources—Chapters 16 and 17.

Informatics in clinical practice—Chapters 18 and 19.

I would like to thank Carol Yeung for her help in preparing this book.

Ronald J. A. Trent
Sydney, December 2006

Contents

Contributors

Jonathan W. Arthur • *Discipline of Medicine, Central Clinical School, University of Sydney and Sydney Bioinformatics, New South Wales, Australia*

Jennifer H. Barrett • *Section of Genetic Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, St James's University Hospital, Leeds, United Kingdom*

Michael A. Black • *Department of Biochemistry, University of Otago, Dunedin, New Zealand*

Paul C. Boutros • *Department of Medical Biophysics, University of Toronto, Ontario, Canada*

Allen K. L. Cheung • *Centre for Virus Research, Westmead Millennium Institute, Westmead, New South Wales, Australia*

Enrico Coiera • *Centre for Health Informatics, University of New South Wales, New South Wales, Australia*

Stuart J. Cordwell • *School of Molecular and Microbial Biosciences, University of Sydney, New South Wales, Australia*

Ben Crossett • *School of Molecular and Microbial Biosciences, University of Sydney, New South Wales, Australia*

Andrew Dubowsky • *Genetic Pathology, Flinders Medical Centre, Bedford Park South Adelaide, Australia*

Alistair V. G. Edwards • *Discipline of Pathology, School of Medical Sciences, University of Sydney, New South Wales, Australia*

Piotr G. Fajer • *Institute of Molecular Biophysics, Department of Biological Sciences, Florida State University, Tallahassee, Florida, USA*

David C. Y. Fung • *School of Information Technologies, Faculty of Science, University of Sydney, New South Wales, Australia*

Scott A. Grist • *Genetic Pathology, Flinders Medical Centre, Bedford Park South Adelaide, Australia*

Christopher A. Haiman • *Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA*

Brett D. Hambly • *Pathology Discipline, Bosch Institute, School of Medical Science, University of Sydney, New South Wales, Australia*

Marcus Hinchcliffe • *Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

Nan Hu • *Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, USA*

ANTHONY M. JOSHUA • *Department of Medical Oncology, Princess Margaret Hospital, Toronto, Canada*

HUONG LE • *Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

MAXWELL P. LEE • *Laboratory of Population Genetics, Center for Cancer Research, NCI, Bethesda, Maryland, USA*

SIAW-TENG LIAW • *School of Rural Health, University of Melbourne, Shepparton, Victoria, Australia*

DONALD R. LOVE • *School of Biological Sciences, University of Auckland, Auckland, New Zealand*

ALEXANDRE MENDES • *Newcastle Bioinformatics Initiative, University of Newcastle, New South Wales, Australia*

PABLO MOSCATO • *Newcastle Bioinformatics Initiative, University of Newcastle, New South Wales, Australia*

CECILY E. OAKLEY • *Institute of Molecular Biophysics, Department of Biological Sciences, Florida State University, Tallahassee, Florida, USA*

CHI N. I. PANG • *Biomolecular Sciences, Faculty of Science, University of New South Wales, Australia*

ANASSUYA RAMACHANDRAN • *Department of Obstetrics and Gynaecology, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand*

PETER SCHATTNER • *Department of General Practice, Monash University, Clayton, Victoria, Australia*

RODNEY J. SCOTT • *Discipline of Medical Genetics, University of Newcastle, New South Wales, Australia*

ANDREW N. SHELLING • *Department of Obstetrics and Gynaecology, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand*

VITALI SINTCHENKO • *Centre for Infectious Diseases and Microbiology-Public Health, Western Clinical School, The University of Sydney, New South Wales, Australia*

BARRY SLOBEDMAN • *Centre for Virus Research, Westmead Millennium Institute, Westmead, New South Wales, Australia*

DANIEL O. STRAM • *Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA*

GRAEME SUTHERS • *Familial Cancer Unit, Children's Youth and Women's Health Service, North Adelaide, Australia*

PHILIP R. TAYLOR • *Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, Bethesda, Maryland, USA*

RONALD J. A. TRENT • *Department of Molecular and Clinical Genetics, University of Sydney Central Clinical School, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

CARL VIRTANEN • *University Health Network, Microarray Centre, Toronto Medical Discovery Tower, Toronto, Ontario, Canada*

MELANIE Y. WHITE • *Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA*

MARC R. WILKINS • *Biomolecular Sciences, Faculty of Science, University of New South Wales, Australia*

JAMES WOODGETT • *Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada*

HOWARD H. YANG • *Laboratory of Population Genetics, Center for Cancer Research, NCI, Bethesda, Maryland, USA*

JEAN YEE HWA YANG • *School of Mathematics and Statistics, University of Sydney, New South Wales, Australia*

BING YU • *University of Sydney Central Clinical School, Department of Molecular and Clinical Genetics, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia*

1

In Silico Gene Discovery

Bing Yu

Summary

Complex diseases can involve the interaction of multiple genes and environmental factors. Discovering these genes is difficult, and *in silico* based strategies can significantly improve their detection. Data mining and automated tracking of new knowledge facilitate locus mapping. At the gene search stage, *in silico* prioritization of candidate genes plays an indispensable role in dealing with linked or associated loci. *In silico* analysis can also differentiate subtle consequences of coding DNA variants and remains the major method to predict functionality for non-coding DNA variants, particularly those in promoter regions.

Key Words: Gene discovery, complex disease, data mining, prediction, data hosting, *in silico* prioritization, haplotype inference, simulation.

Abbreviations: cM – centimorgan; EST – expressed sequence tag; LD – linkage disequilibrium; OMIM – Online Mendelian Inheritance in Man; SNP – single nucleotide polymorphism

1. Introduction

The completion of the Human Genome Project was a milestone in medical science. However, sequencing of approximately 3 billion bases was only the start in deciphering the human genome. Ultimately, the goal of the Human Genome Project is to understand the biology and underlying physiology of human health and disease. Today, common public health issues include cardiovascular disease, stroke, cancer, diabetes, obesity, and mental health problems. These are the major health and economic challenges in many communities

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

(1–3). Compared to simple monogenic or mendelian disorders, these common diseases are "complex," since their etiologies involve many genes and environmental risk factors as well as complicated gene–gene and gene–environment interactions ***(4***,***5)***. Discovery of disease-related genes improves our knowledge of disease etiology and pathogenesis, and subsequently will lead to novel diagnostic and therapeutic methods in treating common diseases.

Gene discovery started in the late 1970s with a *functional* approach, in which the gene product and its function were used to identify a gene ***(4***,***5)***. The discovery of the α-globin gene for the thalassemia syndrome was one such example. However, the identification of abnormal gene products is not always possible in many diseases with a gene defect or with a genetic component. Therefore, *positional* cloning has become an important strategy in gene discovery since the late 1980s ***(5)***. This method bypasses the protein and enables direct cloning of genes on the basis of chromosomal position. It has already proved highly successful for simple monogenic diseases, such as cystic fibrosis, Huntington disease, and many other rare disorders during the past two decades. Most of these gene discoveries were achieved with family-based linkage analyses. However, the latter approach has limited capacity in identifying genes with low penetrance and modest effect, which are traits predicted to be present in most common complex diseases. As an alternative, genetic association studies have become increasingly popular for gene discovery in complex traits. This approach involves the testing of gene sequence variations for their potential involvement in complex diseases. This is usually demonstrated by showing disease alleles that are more or less common in affected individuals than they are in the general population ***(1***,***2)***. In contrast to linkage analysis, association studies have relatively more power to detect modest effects (*see* also **Chapter 2** for further discussion on Mendelian and complex diseases).

In silico gene discovery is a complementary strategy and significantly enhances the likelihood of finding genes, although on its own it is not enough. It integrates biology, computer science, and mathematics to facilitate locus mapping, gene search, and DNA variant identification (**Fig. 1**) ***(5)***. "*In silico*" describes the search for particular information stored in computers, but also includes tasks such as organizing, analyzing, and predicting increasingly complex data arising from molecular and clinical studies with the aid of computers. This chapter focuses on how the *in silico* approach can facilitate gene discovery.

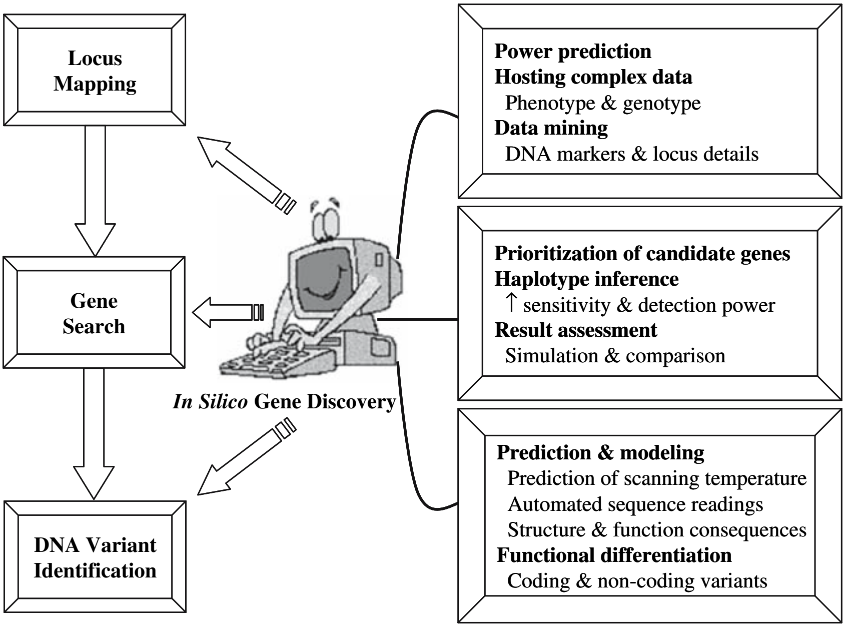

Locus Mapping
Gene Search
DNA Variant Identification
In Silico Gene Discovery
Power prediction
Hosting complex data
Phenotype & genotype
Data mining
DNA markers & locus details
Prioritization of candidate genes
Haplotype inference
↑ sensitivity & detection power
Result assessment
Simulation & comparison
Prediction & modeling
Prediction of scanning temperature
Automated sequence readings
Structure & function consequences
Functional differentiation
Coding & non-coding variants

Fig. 1. Potential contributions of the *in silico* approach to the discovery of complex disease-related genes.

2. Methods

Gene discovery involves three basic steps: (1) locus mapping, (2) gene search, and (3) DNA variant identification (**Fig. 1**). The software resources for *in silico* discovery are extensive in either free or subscription-based forms. Some *in silico* resources are listed in **Tables 1–3**.

2.1. Locus Mapping

2.1.1. Power Prediction

Gene discovery is demanding of time and resources *(5)*. It would be unwise to start this approach if a study did not have sufficient power to obtain a conclusive result. Therefore, power prediction is a useful starting point. *In silico* simulation can be applied to estimate the power of the available collection (either families or case/control cohorts) in the presence of gene effect only, gene–gene interaction, or gene–environment interaction *(6–8)*.

Table 1
***In silico* resources for locus mapping**

Note	Internet Address
List of genetic analysis software	http://www.nslij-genetics.org/soft/
FBAT: Family Based Association Test including power prediction function	http://www.biostat.harvard.edu/~fbat/
QUANTO: a program for power and sample size calculations for genetic epidemiology studies	http://hydra.usc.edu/gxe
Human genome sequences	
National Center for Biotechnology Information (NCBI, GenBank)	http://www.ncbi.nlm.nih.gov/
Ensembl	http://www.ensembl.org/index.html
University of California, Santa Cruz (UCSC) Genome Browser	http://genome.ucsc.edu/
International HapMap Project	http://www.hapmap.org/
GeneSeeker: extract & integrate human disease-related information from many web-based genetic databases	http://www.cmbi.ru.nl/GeneSeeker/
Phenotype resources	
PubMed: the main repository of published biomedical literature	http://www.ncbi.nlm.nih.gov/
OMIN: Online Mendelian Inheritance in Man	Same as above
OMIA: Online Mendelian Inheritance in Animals	Same as above

In a family-based design, the conditional power of available offspring can be calculated based on informative parental genotypes using the FBAT program (**Table 1**). Any association can then be tested in the most promising combinations of DNA markers and phenotypes. This approach also avoids the penalty that comes when multiple comparisons are made.

In a population design, power estimation can be easy when the sample sizes are fixed, and the frequencies of genotypes and/or alleles are known in the target population (*see* **Note 1**) *(8)*. However, it can be complicated if the above information is not available. Ambrosius and colleagues *(9)* developed a program for power calculations based on a Bayesian approach. This program

Table 2
***In silico* resources for gene searching**

Note	Internet Address
STACKdb™ (Sequence Tag Alignment and Consensus Knowledgebase): an EST database of virtual human transcript and is separated by tissue type.	http://www.sanbi.ac.za/Dbases.html
Heart-specific transcripts	http://tcgu.bwh.harvard.edu/
BioManager: a subscription-based bioinformatics workspace that provides a single, user-friendly and intuitive web interface.	http://www.angis.org.au/
PROSPECTR (PRiOritization by Sequence & Phylogenetic Extent of CandidaTe Regions): an alternating decision tree that has been trained to differentiate potential disease-related genes.	http://www.genetics.med.ed.ac.uk/prospectr/
Haplotypes estimation from unphased genotype data in unrelated individuals	
Phase	http://www.stat.washington.edu/stephens/software.html
SNPHAP	http://www-gene.cimr.cam.ac.uk/clayton/software/
Haplo.stat	http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm
Haplotyper	http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm

can test an association between one or more genetic variants and a phenotype of interest, effectively dealing with the sampling variability and allowing for allele frequency uncertainty for both qualitative and quantitative traits. QUANTO is another program (**Table 1**) that can compute either power or required sample size for association studies of genes, environmental factors, gene–environment interactions, or gene–gene interactions ***(6)***. This program can also cope with both qualitative and quantitative outcomes.

Table 3
***In silico* resources for DNA variant identification**

Note	Internet Address
Prediction DNA scanning temperatures using DHPLC	http://insertion.standord.edu/melt.html
Sorting Intolerant From Tolerant (SIFT): predictions for which amino acid substitutions will affect protein function	http://blocks.fhcrc.org/sift/SIFT.html
PolyPhen (Polymorphism Phenotyping): prediction of functional effect of human non-synonymous SNPs.	http://coot.embl.de/PolyPhen/
SNPs3D: a web tool that assigns functional effects of non-synonymous SNPs based on structure and sequence analysis.	http://www.snps3d.org/
Orthologous gene information:	
GenBank	http://www.ncbi.nlm.nih.gov/
Ensembl	http://www.ensembl.org/index.html
Inparanoid	http://inparanoid.cgb.ki.se/
RESCUE-ESE: predict potential splicing elements using a statistical/computational method	http://genes.mit.edu/burgelab/rescue-ese/
Prediction of potential SNP effect at transcriptional level	
PupaSNP Finder	http://pupasnp.bioinfo.ochoa.fib.es/
rSNP_Guide	http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/
Consensus	http://bioweb.pasteur.fr/seqanal/interfaces/consensus.html
MEME (Multiple Em for Motif Elicitation)	http://meme.sdsc.edu/meme/
AlignACE	http://atlas.med.harvard.edu/
BioProspector	http://ai.stanford.edu/∼xsliu/BioProspector/

With reliably estimated power, the investigators can make a more informed decision on whether to proceed with the study. *In silico* simulations are also useful to guide the recruitment of effective numbers of cases and controls (*see* **Note 1**) ***(6)***. The simulation results are needed to convince granting bodies to support a gene discovery project.

2.1.2. Hosting Abundant and Dynamic Information

2.1.2.1. Phenotyping

Clinical assessment or phenotyping of affected and unaffected individuals has to be completed early in gene discovery (**Fig. 1**). Success is very much dependent on the confidence with which phenotypes were assigned from the start. Phenotyping is not necessarily straightforward, and it requires a predefined protocol of inclusion and exclusion criteria along with carefully designed database(s) to host increasingly complex data.

The number of individuals increases significantly from single family members in a linkage study to large cohorts of cases and controls numbering in the hundreds or thousands, as well many families (with variable sizes) in genetic association studies. There are large, heterogeneous and highly interrelated datasets at the initial stage. These data should be collected in a systematic way with predefined protocols. It is common to have repeated measurements in complex traits, such as body mass index in obesity ***(2)***. These data can be important clues in a gene search, and the longitudinal data can also increase the reliability of a phenotype. A careful record of the onset of disease can be useful in the stratification of extreme subsets since an early onset usually suggests a strong genetic component, whereas a late onset might be more likely attributable to an environmental effect. The environmental data can be another crucial component to analyze, especially when gene-environmental interactions play a role in the relevant disease or complex trait. The phenotypic database should have a complete and a uniform set of data encompassing all key aspects of the phenotype. Finally, the re-classification of study subjects is not uncommon in gene discovery, because conventional definitions are revisited or refocused to a particular intermediate phenotype or endophenotype. The phenotypic database should be sufficiently flexible to cope with broad diagnostic criteria as well as changes in the phenotype.

2.1.2.2. Locus Information

With the available sequence data from the Human Genome Project, the *in silico* approach eliminates the laboratory-based contig creation and time-consuming sequence walking. Genomic sequences and linkage disequilibrium

(LD) information within the region of interest can be easily retrieved from the human genome databases and the HapMap project (*see* **Table 1** and **Chapter 3**). It is essential to construct a well-organized locus database with all relevant information such as retrieved sequences, DNA markers, LD blocks, and all available transcripts. This locus database can provide a platform for fine mapping and the subsequent gene isolation as well as DNA variant identification (**Fig. 1**); *see* **Note 2**.

The increasingly larger sizes for complex-trait loci have increased the work required for DNA variant scanning. This has gradually been replaced by the candidate gene approach (see candidate prioritization in section 2.2.1). The obvious requirement for a gene's becoming a candidate is its expression in the disease-related tissue or organ. Therefore, it is quite useful to construct a second transcriptome database for the target tissue or organ. The transcripts in the second database can be classified into constitutive, structural, and functional genes, and incorporated with all potential alternative splicing data. If possible, the microarray expression data from unaffected and affected sources can be integrated into this transcriptome database, which can augment the chance of gene identification. Genes that overlap the locus database and the transcriptome databases should be considered as an essential criterion in gene identification (*see* **Note 2**).

2.1.3. Data Mining

2.1.3.1. Selection of DNA Markers

DNA markers with known chromosomal locations compose another essential component in locus mapping (**Fig. 1**). A plethora of genetic variations are available in the human genome ***(4,5)***. These include the restriction fragment-length polymorphisms (RFLP), short tandem repeats (also called microsatellites), and single nucleotide polymorphisms (SNPs). It is one of the key tasks for *in silico* gene discovery to identify and select a set of appropriate DNA markers.

In linkage analysis, several selected sets of microsatellites, in the hundreds, are available for a genome-wide scan. Since these markers are spaced millions of bases apart, fine mapping is usually needed to narrow down the linked region after the locus is identified. It is important to perform an *in silico* search in the region of interest and to identify all the available DNA markers. High-quality (informativeness) of selected DNA markers can provide the valuable clue in fine mapping in given families and greatly facilitate the identification of the target gene(s).

In silico search is even more critical in association studies using a gene-based candidate approach or fine mapping under a linkage peak, looking for a particular marker or a set of markers to be correlated with disease (or trait values) across a population rather than within families. Because segments of LD are measured in a few thousands to tens of thousands of bases, a large number of DNA markers are required to scan a candidate region ***(10)***. Mining of valid and informative SNPs is important to initiate an association study. *In silico* tools can increase the efficiency of marker selection via simulation ***(7)*** in order to capture more genetic variation in candidate genes and to avoid multiple comparisons. If SNPs are scarce in one candidate region, both *in silico* and a more traditional experimental approach may be required. If there are abundant SNPs in another region, an *in silico* search can be helpful to reduce redundancy. With the regional LD information, the *in silico* approach can define a set of SNPs, forming a haplotype in a particular candidate region, which would essentially "tag" other variants not directly tested ***(11)***. This makes it possible to survey substantial fractions of human variation in a cost-effective manner. It can be assumed that the causal DNA variants or trait-modifying alleles will be in strong LD with one of the tagging SNPs being directly genotyped if the set of SNPs are sufficiently dense (*see* also **Chapter 3**).

2.1.3.2. Retrieving Available Genes

In silico search plays a role in the annotation of genes in an identified locus. As discussed previously, the target tissue-expressed and locus-specific genes can be retrieved through the database overlap (*see* **Note 2**). Some of these retrieved genes have known biological functions while others do not. It is crucial to be as inclusive as possible in positional cloning. The potential functions of these unknown genes can be predicted through sequence alignment with available known genes in different species. The presence of characteristic motifs is helpful for determining putative function via similarity searches. Alternatively, the othologous block i.e. the synteny region, that is assumed to be derived from the immediate ancestor of two species can be detected in model organisms using sequence alignment between a pair of genomes ***(12)***. Then the potential function of a gene or a locus can be studied in model organisms. Data mining from available microarray studies, tissue-2D gel or two-hybrid systems are also useful to identify the potential functions of unknown genes, especially for the novel ones without any identified sequence similarity.

2.1.3.3. Tracking New Knowledge

Gene discovery is an ongoing process. It is necessary to combine observations and existing knowledge at the time of defining candidate genes. Proficiency in computer-aided literature search would identify new knowledge that researchers may not always be aware. Equally important is to keep abreast of emerging knowledge. Online tools are available to support literature-based discovery in the life sciences ***(13)***. GeneSeeker is one such tool. It gathers information that conforms to investigator-defined criteria from several databases such as OMIM, human and mouse genomes and expression data ***(14)*** (*see* **Table 1**). Other similar tools are available based on sequence or protein motif similarity and Gene Ontology (GO) terms ***(15–17)***. Automated literature tracking is also useful to guide the prioritizing process for candidate genes (*see* **section 2.2.1**).

2.2. Gene Search

This is a time-consuming task **(Fig. 1)**. *In silico* gene discovery can accelerate the process or bypass many steps in the traditional gene discovery process.

2.2.1. Candidate Gene Prioritization

Prioritizing an increasing number of candidate genes for DNA variant screening is a challenge. This is particularly so when the region of interest, especially for complex diseases, can expand from a few centimorgans (cM) to 10–40 cM ***(18)***. It is still too laborious and expensive to genotype thousands to a few hundred thousand SNPs or to sequence all genes in a region of interest in complex diseases. Prioritization of candidate genes is cost-effective, and can maximize the chance of success.

2.2.1.1. Common Practice in Gene Prioritization

Usually candidate genes are prioritized through an integration of various datasets including phenotypic and/or expression data, linkage results in animal models, knowledge of biologic pathways, and genes in the chromosomal region(s) of interest. This is done to match the functional annotation of particular gene(s) to knowledge of the disease or phenotype in question, and to exclude any candidate genes that have not fulfilled certain criteria. The *in silico* approach can facilitate this process (*see* **Note 2**). However, the standard prioritization approach has limitations. The link between a gene and the phenotype of interest tends to be weak in complex traits or disease. It is not always possible to identify an apparent matching between the gene's function and the related phenotype. Target genes can be embedded in genes with unknown function or can have an unexpected connection ***(1)***. Functional annotation of the human genome is incomplete and biased toward better-studied genes. The annotation is time-consuming and unavoidably error-prone ***(19)***. New bioinformatics tools have now been developed for complementing common practice in the prioritization process.

2.2.1.2. In Silico Tools in Gene Prioritization

Adie and co-workers ***(18)*** developed a unique program, “Prospectr,” which is independent on any available functional annotation (**Table 2**). This program has integrated many sequence-based features that have been derived from detailed studies of >1,000 disease genes available from OMIM. The features used in building the alternating decision tree include the numbers of exons and CpG islands, the length of the 3’ untranslated region, the distance to the nearest neighboring gene, the tissue-specific expression, and the sequence conservation across species. As the Prospectr outputs are the alternating decision tree scores, it can adjust the specificity (precision) at the expense of sensitivity (recall). This program can help prioritization involving large regions of interest in minutes, and select candidate genes for further case/control association or DNA variant identification (**Fig. 1**).

2.2.1.3. Gene Prioritization from Model Organisms

Studying the relevant disease in animals is another alternative to assist in prioritizing the candidate genes since this type of study can be conducted in a carefully controlled environment. Although results from animal studies cannot unconditionally translate into data of relevance to humans, mechanistic insights can help improve experimental design in the human studies. This is particularly interesting in animals with evolutionarily related pathways and conserved gene sequence and function. For example, a quantitative trait locus for coronary artery disease (atherosclerosis 1) was identified on mouse chromosome 1 that renders C57BL16 mice susceptible to diet-induced atherosclerosis. *Tnfsf4* (tumour necrosis factor superfamily, member 4) was then identified as the underlying gene. This leads to a link between coronary artery disease and immunological pathways in humans [for review see ***(3)***]. *In silico* search can also provide the clue in identifying relevant animal models from Online Mendelian Inheritance in Animals (OMIA, **Table 1**) or published literature with links to potential human homologs. It would be worthwhile knowing if a gene in a disease-related animal is in a syntenic region to the identified human locus. With the help of whole genome information from at least 25 model organisms (listed in GenBank and Ensembl, **Table 1**), the use of comparative genomics can accelerate gene discovery and identify the promising gene candidates for common complex diseases.

2.2.2. Contribution to Analysis

2.2.2.1. In Silico Haplotype Inference

The prioritized candidate genes have to be tested experimentally. Single SNPs have limited informativeness. The achievement of the HapMap project provides a complete reference panel of LD structure in the human genome (*see* **Chapter 3**). It leads to the natural progression from a single marker analyses toward multimarker haplotype analyses, especially in gene-based association studies or fine mapping. Haplotypes refer to a set of SNP alleles that are co-segregate on a single chromosome. Haplotype tests are more efficient with significantly improved statistical power in association studies compared with the pairwise analysis between single SNPs and a phenotype ***(10,20)***. Different combinations of particular SNP alleles in the same gene may act as meta-alleles and significantly increase the sensitivity in phenotype-genotype correlations. If the most important variants for complex diseases are non-coding sequence, haplotyping is the most cost-effective screening method. "Diplotype" represent the interactions between various haplotypes that come together in an individual's diploid genome and provide more power to discriminate subtle effects.

Empirically, haplotypes, i.e., the phase of each DNA marker, are determined based on the genotype data from parents and grandparents or sequencing after cloning of long-range PCR products. The familial data are difficult or impossible to collect especially in late onset diseases. The experimental methods of long-range amplification, cloning, and sequencing are expensive and labor-intensive. In order to carry out high-throughput haplotype analysis, *in silico* inference is essential in dealing with SNP genotype data with ambiguous phase for unrelated individuals. Some available methods for haplotype prediction include Phase, SNPHAP, Haplo.stat, and Haplotyper (**Table 2**). These programs use different algorithms in haplotype inference, which include the expectation-maximization algorithm, Bayesian method and parsimony or subtraction method [for review *see* ***(21)***]. Compared with the empirical results, most of the above programs can assign about 90% correct haplotypes for individuals heterozygous for up to three SNPs and are about 80% correct for up to five heterozygous sites. These methods can identify every haplotype with a frequency above 1%. Incorrect haplotypes are possible but if so they will not exceed a 1% frequency ***(21)***. It is also useful to estimate haplotypes using several methods. Consistency in the estimation results would imply the approximation is more likely to be correct. Inferred haplotypes with their probabilities greatly enhance the capacity to confirm or exclude a target candidate gene.

2.2.2.2. Simulation in the Assessment of Association Results

False positive results in association studies remain a major concern ***(4)***. One of the underlying contributors to this is a chance finding due to multiple testing. One way to address this is applying *in silico* power. Datasets can be simulated according to the null hypothesis in the linkage or association study using available pedigree structures, case-control collection, marker map or LD information, allele frequencies, and patterns of missing data as in the study itself. Observed significance levels can then be assessed. A permutation procedure can also test how often a random subset of the same size yields a p value equal or smaller to the one observed ***(22)***. With advances in computer technology, simulation has become the standard method of assessing the significance of association studies in gene discovery.

2.3. DNA Variant Identification

Once a candidate gene is isolated, it becomes necessary to show the gene has a causal relationship with the disease of interest. This usually involves extensive DNA scanning to search for the disease-causing (pathogenic) mutation(s) or modifying (either susceptible or protective) DNA variant(s) (**Fig. 1**). Although extensive scanning is daunting, it is the ultimate goal for gene discovery to uncover the molecular etiology, and to identify potential targets for disease prevention and/or intervention. The *in silico* approach greatly enhances the capacity to characterize functional DNA variants.

2.3.1. In Silico Support for DNA Scanning

High throughput and effective methods such as DHPLC are essential in DNA scanning with large linked or associated loci. A predictive algorithm is available to assist in the amplicon design and to suggest screening temperature(s) in DHPLC analysis (*see* **Table 3**) so that DNA scanning can achieve high sensitivity. In the identification of DNA variant(s) with a sequencing approach, software is available for computer-assisted DNA sequence reading (*see* **Chapter 10**). The latter computational support markedly relieves the labor-intensive reading process, and increases the sensitivity in DNA variant identification, especially heterozygous changes, from long stretches of normal sequence. However, the *in silico* approach goes far beyond the above supporting role, and it also plays a critical role in prediction of structural and functional consequences of a particular DNA variant.

2.3.2. In Silico Prediction of Mutational Effects

In simple monogenic disorders, DNA variants such as missense mutations, small insertion/deletion (indel), nonsense changes, and altered splicing sites can have profound effects on the structure and function of the corresponding proteins, with major functional consequences. Whether a particular amino acid is conserved can be identified through the multiple sequence alignment with the orthologous sequences from different species ***(12)***. The impact of a missense mutation can be modeled if the protein or motif structure is available in the mutation region ***(23)***. Nonsense mutations and frame-shift changes can introduce an early termination and lead to nonsense-mediated decay or a truncated protein with obvious functional changes (*see* **Chapter 11**).

2.3.3. In Silico Differentiation of Functional DNA Variants

In complex diseases, the tested SNP may be a true variant that leads directly to disease susceptibility. Alternatively it may merely be in LD with an unobserved susceptibility allele i.e. functional variant. It is important to differentiate the functional DNA variants from numerous neutral DNA changes. Application of *in silico* methods can provide some answers in a fraction of the time that it once took in the laboratory.

2.3.3.1. Visualizing the Variant Location

DNA variants can be found anywhere near or in a gene, and their potential influence on the function of DNA, RNA, and proteins is closely related to the genomic location. Many *in silico* resources are available in assist in the visualization of variant loci and their annotation. In the public domain there are UCSC Genome Browser and Ensembl. Both are quite useful with a friendly interface (*see* **Table 1** and **Chapter 17**).

2.3.3.2. Identifying Potential Subtle Consequences of Coding DNA Variants

Not all non-synonymous DNA variants have obvious deleterious effects on the structure or function. These changes in a complex trait may have subtle effects on protein–protein interactions, a fine balance of different alternative spliced isoforms, glycosylation site changes, or may make a protein function less optimal.

The *in silico* approach can help to differentiate a functional variant from a polymorphic non-synonymous change by taking many factors into consideration ***(24)***. These factors includes the protein's structural properties such as solvent accessibility, location within β-strands or active sites, and participation in disulfide bonds; and altering protein stability such as ligand binding, catalysis, allosteric regulations, and post-translational modification. The SIFT and Polyphen programs (**Table 3**) are commonly used resources for classifying uncharacterized non-synonymous SNPs ***(25)***. These programs can estimate a particular amino acid substitution as tolerant or damaging based on the position in homologous genes. It can assess local functionality using both protein structural and evolutionary information that are obtained by comparing orthologous genes. A wide variety of orthologs can be retrieved from many databases such as GenBank, Ensembl or Inparanoid (**Table 3**). SNPs3D is another resource for inferring the molecular function of DNA variants based on structure and sequence analysis.

Synonymous changes may not necessarily be "silent" as previously thought. These changes can be part of exonic splicing regulatory elements that affect the gene expression via altered splicing selection or efficiency. For example, a synonymous change (TTC to TTT, Phe to Phe) was identified in exon 7 of the Survival of Motor Neuron 2 (*SMN2*) gene. This change disrupts an exonic splicing enhancer site and creates an exonic splicing repressor site ***(26,27)***, which is enough to prevent efficient exon 7 splicing. In the absence of the *SMN1* gene, this C→T transition is directly related to the severity of spinal muscular atrophy. Since these functional synonymous changes are under selective pressure, they can be predicted or identified through multiple sequence alignment of the orthologous genes in closely related mammalian genomes ***(12)***. One program, RESCUE-ESE, has been developed to predict potential exonic splicing elements using a statistical/computational method ***(28)*** (**Table 3**). This program can effectively predict the loss of the exonic splicing enhancer site of the above synonymous change in the *SMN2* gene. This type of *in silico* approach greatly enhances the capacity for differentiating exonic variants (either synonymous or non-synonymous) as putative exonic splicing elements from non-functional ones.

2.3.3.3. Predicting Potential Functionality of Non-Coding DNA Variants

DNA variants can be located in the non-coding regions, including the promoter region, introns, the 5'- and 3'-untranslated regions, or the polyadenylation signal site. Some can be random polymorphisms with neutral effects; others can have an impact on a disease process via their influence on gene expression. It has been reported that introns can contain regulatory elements for gene expression and be significantly associated with increased risk for complex disease such as diabetes ***(1,29,30)***. Assessment of non-coding DNA variants apart from those in well-conserved splicing sites remains a challenge since it is difficult to separate regulatory variation (*cis*-acting and *trans*-acting factors) from the cellular environment.

Regions that are biologically important tend to be more constrained by evolution and therefore more conserved. *In silico* analysis provides a unique platform for identifying the promoter DNA variants at the conserved sites and potential binding motifs for transcription factors ***(31,32)***. This remains the only way of trying to predict if a DNA variant will affect expression levels. PupaSNP Finder is one such tool for identifying SNPs that could have an effect on transcription (*see* **Table 3**). rSNP_Guide contains annotations of SNPs based on potential effects on regulation. Other *in silico* tools for predicting transcription factor binding sites include Consensus, MEME, AlignACE, and BioProspector (*see* **Table 3**).

Over the past few years, *in silico* analysis of alternative splicing has emerged as an important new focus. It is of great interest that each gene can be "reused" to create multiple functions and new modes of regulation. Potential non-coding DNA changes can alter the essential recognition sites for the spliceosome, intronic splicing enhancers or silencers, and the binding sites for splicing regulatory factors. It is almost impossible to recognize the above variants without *in silico* analysis. The *in silico* identification of DNA variants that affect splicing ***(28)*** and deciphering the regulatory control mechanisms that govern gene expression would simplify interpretation of complicated and puzzling data in the complex diseases.

3. Conclusions

Genetic association studies currently focus on a candidate gene or a candidate region. These studies have an inherent disadvantage because of limited knowledge of the molecular basis for complex diseases. Ultimately, with the increased throughput and reduced costs of genotyping, genome-wide association studies based on LD-directed tag-SNPs will become feasible and allow the identification of the genetic contributions in complex diseases to be made more efficiently. This genome-wide approach will heavily rely on *in silico* strategies.

The integration of the *in silico* approach with the genome-wide association studies will provide enormous scope for testing genetic effects or interactions across genetic regions. Such integration will lead the genetics of complex diseases to a point of success comparable to where mendelian genetics now firmly resides.

4. Notes

1. As discussed in 2.1.1, the power of a case-control study can be estimated when the sample sizes are fixed and the frequencies of genotypes and/or alleles are known in the target population. For example, as an exercise let us determine whether the case-control study reported by Gayagay et al. ***(33)*** has enough power to detect a frequency increase of ACE allele I (angiotensin II-converting enzyme gene insertion/deletion polymorphism) from 0.43 in the unmatched controls (R) to 0.57 in the cases (C), i.e., $p_R = 0.43$ and $p_C = 0.57$.

   The number of alleles as reported were 128 and 236 for the case and control groups, respectively, i.e., $n_C = 128$ and $n_R = 236$. The geometric mean of the alleles being tested can be calculated as follows:

$$n = \frac{2 \times n_C \times n_R}{n_C + n_R} = 166$$

   The effect size index (h) can be calculated using the allele frequencies from the cases and controls after the nonlinear transformation. As suggested by Lalouel and Rohrwasser ***(8)***, using the arcsin transformation $\theta = 2\arcsin\sqrt{p}$ calculate θ_C and θ_R and h = θ_C - θ_R. The calculator should be set to use radian rather than degree in arcsin transformation.

$$h = \theta_C - \theta_R = 2\arcsin\left(\sqrt{p_C}\right) - 2\arcsin\left(\sqrt{p_R}\right) = 0.28$$

   To calculate power at a significance level of α = 0.05, $z_{1-\alpha}$ can be obtained from a table of normal distributions (**Table 4**). When $z_{1-\alpha} = 1.65$, it gives the value of 0.9505, i.e., the value of t, in the table, for which the shaded area under the curve is equal to 0.95. The power of the study can be calculated as follows:

$$z_{1-\beta} = h\sqrt{\frac{n}{2}} - Z_{1-\alpha} = 0.90$$

   This gives the value to be referred to in the table of the normal distribution (**Table 4**), i.e., z = 0.90 ⟶ 0.8159. It means the power 1-β = 0.82. Let us further assume that we wanted to increase the power to 0.85, but we could not increase the case numbers. It would still be possible to achieve a power of 0.85 by increasing the control alleles from 236 to 384, i.e., case-control ratio of 1:3.
   If we wanted to estimate the genetic relative risk (or odds ratio) detectable with a given number of cases, the Quanto program (**Table 1**) could be used. Let us go back to the Gayagay's study again as an example. First assume ACE I allele frequency of 0.43 and the log-additive as the mode of inheritance (gene effect: G = 2 if II, G = 1 if ID and G = 0 if DD). The calculated genotypes frequencies would be II = 0.18, ID = 0.49, and DD = 0.32, and the observed genotype were

Table 4
Simplified normal distribution table (Probability content from 0 to z)

z	0	0.01	0.02	0.03	0.04	0.05	…	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	…	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	…	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	…	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	…	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	…	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	…	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	…	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	…	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	***0.7995***	0.8023	…	0.8133
0.9	***0.8159***	0.8186	0.8212	0.8238	0.8264	0.8289	…	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	…	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	…	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	…	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	…	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	…	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	…	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	***0.9505***	…	0.9545
…	…	…	…	…	…	…	…	…
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	…	0.9990

II = 0.18, ID = 0.51, and DD = 0.32 as reported ***(33)***, in which there is no significant difference. Therefore, this polymorphism is in Hardy–Weinberg equilibrium. Finally, let us assume that there would be 64 cases with 128 unmatched controls. As we see from the results in **Table 5**, the case numbers of 60 and 70 have the power of 0.78 and 0.84, respectively. Hence, the case-control study from Gayagay et al. (case number of 64) would have a power of 0.80 to detect the genetic relative risk of 1.85 and above.

2. In a recently completed gene-based association study, our aim was to identify the complex trait gene(s) related to cardiovascular performance in elite endurance athletes ***(34)***. Genome-wide scans based on the maximum oxygen uptake (VO_2max) as the endurance phenotype identified four loci related to baseline VO_2max and five loci in response to endurance training ***(35)***. Although these linked loci were only suggestive, we started the gene discovery steps assisted by an *in silico* approach (**Fig. 2**). The locus on chromosome 2p16.1 was anchored by DNA marker D2S2739. First a locus-specific genomic sequence database (see discussion in section 2.1.2.2) was constructed extending 5 Mb on either side of D2S2739. Expressed sequence tags (ESTs) related to cardiovascular function were collected

Table 5
Estimated results of genetic relative risk using the QUANTO program

Relative Genetic Risk	Number of Cases*	Power**
	50	0.48
1.60	60	0.55
	70	0.62
	50	0.70
1.85	60	0.78
	70	0.84
	50	0.85
2.10	60	0.90
	70	0.94

*The number of cases required for the desired power with unmatched case-control (1:2).
**Power is estimated with the hypothesis of gene effect only and under the log-additive model of inheritance.

from many available databases including the STACK database *(36)* and the Cardiovascular Gene Unit website (heart-specific-transcripts—**Table 2**). These ESTs were used to construct the target-specific expression database (see discussion in section 2.1.2.2). Both locus- and target expressed-databases were hosted in BioManager, a platform provided by the Australian National Genomic Information Service (**Table 2**).

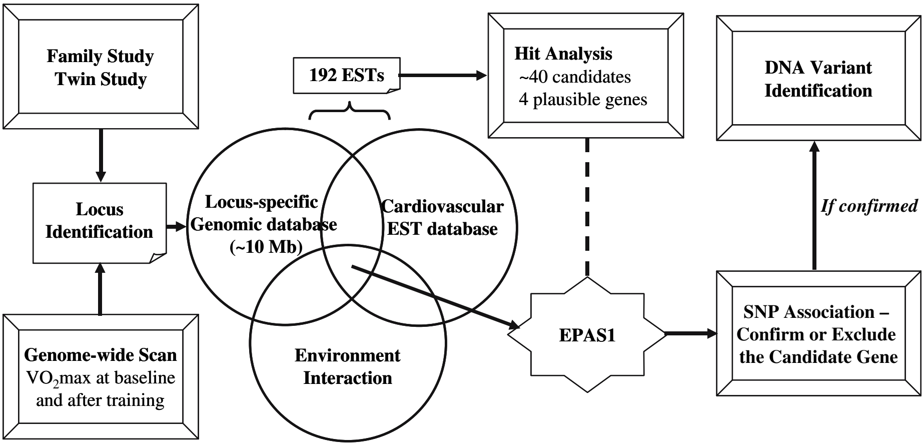


Fig. 2. *In silico* discovery of cardiovascular genes related to athletic performance.

The entry criteria for the candidate genes were then set up, i.e., any candidate gene must be both locus-specific and target-specific. This was achieved by identifying the overlap candidates that were present in both the locus-specific genomic sequence database and the target-specific expression database (**Fig. 2**). One hundred ninety-two ESTs fulfilled the entry criterion. After the extensive sequence analysis and full-length cDNA assembly from multiple EST hits, 40 putative genes were identified. Nearly half of the genes were unknown at the time, and the remainder was scrutinized for known functions. After the exclusion of housekeeping genes, four plausible candidate genes were short listed including calmodulin 2, epithelial PAS protein 1 (*EPAS1* or hypoxia-inducible factor 2 alpha), cytochrome c oxidase subunit VIIa polypeptide 2, and solute carrier family 8 member 1. With further enquiry, only *EPAS1* remained since it can interact with the environment and execute its transcriptional regulation after sensing oxygen deficiency, which occurs commonly during endurance training or competition (**Fig. 2**). Following this, a genetic association study was performed involving 492 cases (elite athletes) and 444 matched controls and 12 selected SNPs. This confirmed that *EPAS1* is at least one of the relevant genes in the 2p16.1 locus ***(34)***.

Acknowledgments

This work was supported by an Australian Research Council Discovery Grant DP0452019.

References

1. Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet*. **38**, 320–323.
2. Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., et al. (2006) A common genetic variant is associated with adult and childhood obesity. *Science* **312**, 279–283.
3. Watkins, H., and Farrall, M. (2006) Genetic susceptibility to coronary artery disease: from promise to progress. *Nat. Rev. Genet*. **7**, 163–173.
4. Thomson, G. (2001) Significance levels in genome scans. *Adv. Genet*. **42**, 475–486.
5. Trent, R. J. (2005) *Molecular Medicine*. Elsevier Academic Press, San Francisco.
6. Gauderman, W. J. (2002) Sample size requirements for matched case-control studies of gene–environment interaction. *Stat. Med*. **21**, 35–50.
7. Laird, N. M., and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet*. **7**, 385–394.
8. Lalouel, J. M., and Rohrwasser A. (2002) Power and replication in case-control studies. *Am. J. Hypertens*. **15**, 201–205.
9. Ambrosius, W. T., Lange, E. M., and Langefeld, C. D. (2004) Power for genetic association studies with random allele frequencies and genotype distributions. *Am. J. Hum. Genet*. **74**, 683–693.

10. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144.
11. de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223.
12. Batzoglou, S. (2005) The many faces of sequence alignment. *Brief Bioinform.* **6**, 6–22.
13. Weeber, M., Kors, J. A., and Mons, B. (2005) Online tools to support literature-based discovery in the life sciences. *Brief Bioinform.* **6**, 277–286.
14. van Driel, M. A., Cuelenaere, K., Kemmeren, P. P., Leunissen, J. A., Brunner, H. G., and Vriend, G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.* **33**, W758–W761.
15. Freudenberg, J., and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18**, S110–S115.
16. Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319.
17. Turner, F. S., Clutterbuck, D. R., and Semple, C. A. M. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* **4**, R75.
18. Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55.
19. Devos, D., and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431.
20. Judson, R., Stephens, J. C., and Windemuth, A. (2000) The predictive power of haplotypes in clinical response. *Pharmacogenomics* **1**, 15–26.
21. Adkins, R. M. (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genet.* **5**, 22.
22. Van Den Bogaert, A., Schumacher, J., Schulze, T. G., Otte, A. C., Ohlraun, S., Kovalenko, S., et al. (2003) The DTNBP1 (dysbindin) gene contributes to schizophrenia, depending on family history of the disease. *Am. J. Hum. Genet.* **73**, 1438–1443.
23. Yu, B. (2004) What is the value of mutation identification in familial hypertrophic cardiomyopathy? *IUBMB Life* **56**, 281–283.
24. Mooney, S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform.* **6**, 44–56.
25. Ng, P. C., and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814.
26. Cartegni, L., and Krainer, A. R. (2002) Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet.* **30**, 377–384.

27. Kashima, T., and Manley, J. L. (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.* **34**, 460–463.
28. Fairbrother, W. G., Yeh, R. F., Sharp, P. A., and Burge, C. B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013.
29. Amador, M. L., Oppenheimer, D., Perea, S., Maitra, A., Cusat, I. G., Iacobuzio-Donahue, C., et al. (2004) An epidermal growth factor receptor intron 1 polymorphism mediates response to epidermal growth factor receptor inhibitors. *Cancer Res.* **64,** 9139–9143.
30. Tokuhiro, S., Yamada, R., Chang, X., Suzuk, I. A., Kochi, Y., Sawada, T., et al. (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.* **35**, 341–348.
31. Bulyk, M. L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201.
32. Pavesi, G., Mauri, G., and Pesole, G. (2004) *In silico* representation and discovery of transcription factor binding sites. *Brief Bioinform.* **5**, 217–236.
33. Gayagay, G., Yu, B., Hambly, B., Boston, T., Hahn, A., Celermajer, D. S., et al. (1998) Elite endurance athletes and the ACE I allele—the role of genes in athletic performance. *Hum. Genet.* **103**, 48–50.
34. Henderson, J., Withford-Cave, J. M., Duffy, D. L., Cole, S. J., Sawyer, N. A., Gulbin, J. P., et al. (2005) The EPAS1 gene influences the aerobic–anaerobic contribution in elite endurance athletes. *Hum. Genet.* **118**, 416–423.
35. Bouchard, C., Rankinen, T., Chagnon, Y. C., Rice, T., Perusse, L., Gagnon, J., et al. (2000) Genomic scan for maximal oxygen uptake and its response to training in the HERITAGE Family Study. *J. Appl. Physiol.* **88**, 551–559.
36. Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., et al. (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9**, 1143–1155.

2

Whole Genome-Wide Association Study Using Affymetrix SNP Chip: A Two-Stage Sequential Selection Method to Identify Genes That Increase the Risk of Developing Complex Diseases

Howard H. Yang, Nan Hu, Philip R. Taylor, and Maxwell P. Lee

Summary

Whole-genome association studies of complex diseases hold great promise to identify systematically genetic loci that influence one's risk of developing these diseases. However, the polygenic nature of the complex diseases and genetic interactions among the genes pose significant challenge in both experimental design and data analysis. High-density genotype data make it possible to identify most of the genetic loci that may be involved in the etiology. On the other hand, utilizing large number of statistic tests could lead to false positives if the tests are not adequately adjusted. In this paper, we discuss a two-stage method that sequentially applies a generalized linear model (GLM) and principal components analysis (PCA) to identify genetic loci that jointly determine the likelihood of developing disease. The method was applied to a pilot case-control study of esophageal squamous cell carcinoma (ESCC) that included 50 ESCC patients and 50 neighborhood-matched controls. Genotype data were determined by using the Affymetrix 10K SNP chip. We will discuss some of the special considerations that are important to the proper interpretation of whole genome-wide association studies, which include multiple comparisons, epistatic interaction among multiple genetic loci, and generalization of predictive models.

Key Words: whole-genome association study, SNP, SNP chip, genotyping, complex disease, genetic interaction, esophageal squamous cell carcinoma (ESCC), generalized linear model (GLM), principal components analysis (PCA).

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

Abbreviations: ESCC – esophageal squamous cell carcinoma; GLM – generalized linear model; HWE – Hardy Weinberg equilibrium; LD – linkage disequilibrium; PCA – principal components analysis; PC1 – first principal component; SNP – single nucleotide polymorphism

1. Introduction

Human diseases are generally classified into two categories: Mendelian disease versus complex disease. These two classes of disease have many distinct genetic and phenotypic characteristics ***(1)***. Mendelian diseases are primarily determined by single genes. They are usually rare and occur in a family setting. Mendelian genes can be mapped by linkage analysis and are identified through positional cloning. Causative mutations are often located within the conserved regions of the affected gene, and these mutations usually change the function of the protein ***(2)***. In contrast, complex diseases are caused by the combined effect of many genes, each of which has a small to moderate effect. Complex diseases are usually common, and examples include cardiovascular diseases, cancer, and diabetes. The complex nature of interactions among the multiple genes and between genes and environmental factors imply that a single locus is unlikely to have enough effect on the risk of the disease. Thus, the linkage analysis approach is less effective. Attention has now been shifted to strategies such as association or linkage disequilibrium (LD) studies that attempt to identify an association between a genetic marker and a disease susceptibility locus ***(3,4)***.

LD can be generated by mutation, migration, selection, and genetic drift. However, LD begins to decay once it is generated. LD between unlinked loci decays rapidly. The rate of LD decay slows down when the two loci are linked. When the two loci are tightly linked, LD can persist through many generations. It is this type of LD that is useful for identifying disease genes. However, a spurious association can exist due to population admixture, sample selection bias, and LD generated from a recent event. So care needs to be taken to reduce the spurious association due to unlinked loci by experimental design and data analysis. Association due to linked loci is more powerful than linkage analysis since such linkage disequilibrium is restricted to small regions of the genome, usually a few kb to 50–60 kb, depending on genetic loci and study population. These regions are also referred to as haplotype or LD blocks.

With the completion of the human genome sequence and the availability of high-throughput genotype technologies, genome wide association studies hold great promise for systematically identifying genetic loci that determine the etiology of complex diseases ***(5)***. Current efforts directed to the identification of disease-causing genes have now shifted from Mendelian diseases to complex diseases. Although the search of Mendelian genes has always focused predominantly on mutations that result from non-synonymous substitution of amino acids, this does not have to be the case for SNPs that affect complex diseases.

Single nucleotide polymorphisms (SNPs) may affect complex diseases through their effects on gene expression, and some examples of this are provided in several recent studies ***(6–8)***. Currently, there are more than 12 million SNPs deposited in GenBank (http://www.ncbi.nih.gov/SNP/), 6.5 million of which have been validated. To facilitate gene discovery, the HapMap project was initiated to expedite the search for the genes that predispose individuals to complex diseases ***(9,10)***. The genetic resources available from the HapMap project provide information on allele frequency, Hardy–Weinberg equilibrium (HWE), linkage disequilibrium (LD), and haplotype structure. They are useful for selecting TagSNPs for association studies ***(11)***. Chapter 3 provides further information on the use of HapMap as a resource.

Esophageal squamous cell carcinoma (ESCC) is one of the most common malignancies in the Chinese population ***(12)***. ESCC showed familial aggregation in the high-risk regions in northern China, indicating that a genetic influence plays a role in the etiology of this cancer. Epidemiology studies suggest that ESCC is a complex disease caused by multiple genetic loci. We previously published a pilot ESCC case-control genome-wide association study using the Affymetrix 10K SNP array ***(13)***. Here, we describe the experimental design and statistical analysis that are important for a genome-wide association study. We describe some of the special considerations concerning the proper interpretation of a whole genome-wide association study, which include multiple comparisons, gene–gene interactions, and generalization of predictive models.

2. Experimental Design and Protocol

Patients and controls in this pilot case-control study of ESCC were described previously ***(13)***. We had 50 ESCC patients and 50 neighborhood-matched controls. Age-, sex-, and neighborhood-matched controls were selected and evaluated within 6 months of the case being diagnosed. The “neighborhood” in China refers to the residence blocks within communities. All individuals and their ancestors lived in Shanxi Province. These individuals were selected to ensure a more homogeneous population structure. Both gender and age are known to affect ESCC. Since our primary goal was to identify genetic risk factors, we selected only male individuals with matched age to remove these confounding factors. The other potential confounding factors included diet, smoking, and alcohol use. Therefore, these were included in the generalized linear model (GLM) as covariates. We found little evidence for any effect due to diet or smoking for most of the SNPs in GLM analyses. These procedures in sample selection and data analyses aim to reduce spurious associations due to sample bias or non-genetic confounding factors so that we can enrich the linkage disequilibrium due to a linked locus. More discussion can be found below (*see* 3. Data Analysis and **Note 1**).

The Affymetrix 10K SNP chip (Affymetrix GeneChip® Mapping 10K Array Set) was designed for simultaneous typing of 11,555 SNPs in the human genome ***(14)***. The mean distance between SNPs is 210 kb and the average heterozygosity for these SNPs is 0.37. The genotype call is determined by the relative intensity from the two alleles, designated as A allele and B allele. More recently, Affymetrix released higher-density SNP chips ***(15)*** including a 100K chip and a 500K SNP chip (Affymetrix GeneChip® Mapping 100K and 500K Array Sets respectively). The mean distance between SNPs is 23.6 kb and 5.8 kb for the 100K SNP chip and 500K SNP chip, respectively. The average heterozygosity of the SNPs for both 100K and 500K chips is 0.30. Other high-throughput genotyping platforms include Illumina Sentrix® Human-1 (109K), HumanHap300 (317K), and HumanHap550 (555K) BeadChips ***(16,17)***. *See* more discussion in **Note 2**.

The genotyping experiment using the Affymetrix 10K SNP chip was described previously ***(13)***. Our more recent genotype data have been generated on higher density SNP chips including Affymetrix 100K and 500K SNP chips.

3. Data Analysis

The 10K SNP chip experiment generated 11,555 genotype data for each sample. We removed 1,291 SNPs because they failed in one of the following quality control steps. The SNP

1. could not be mapped to the NCBI human genome assembly,
2. was homozygous in all cases or all controls, or
3. deviated from Hardy-Weinberg equilibrium (HWE) in the controls. *see* more discussion on HWE criterion in **Note 3**.

Our data analysis strategy is illustrated in **Fig. 1**. We developed a two-stage sequential selection protocol to identify systematically genetic loci that influence an individual's risk of developing disease. A two-stage method was initially suggested as a more cost effective approach for the genomic screen ***(18)***. In stage I, a large number of markers was genotyped on a subset of the samples. In stage II, additional markers in the interesting regions spanning

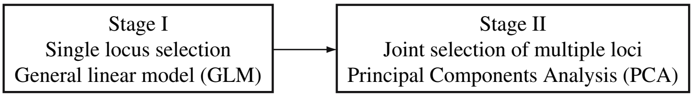

Stage I
Single locus selection
General linear model (GLM)
Stage II
Joint selection of multiple loci
Principal Components Analysis (PCA)

Fig. 1. Two-stage selection method for a whole genome association study. The details of the two-stage method are described in the main text.

the markers selected from stage I exceeding a certain predefined significance level were genotyped on the expanded set of the samples. Recently, the two-stage method was applied to a whole-genome association study ***(19)*** and it was proposed to have greater power to detect two-locus gene interactions that influence complex diseases ***(20)***. More discussion on the two-stage method can be found in **Note 4**.

Our two-stage method extends to multi-locus gene interactions. The strategy is based on the premise that complex disease is caused by combining quantitative effects from multiple genetic loci, each of which has a small effect, but jointly they can account for a significant portion of the risk factors for the disease ***(13)***. The stage I step intends to identify each genetic locus that may contribute to the genetic etiology of the disease. We used a GLM to identify SNPs that may affect disease. With the GLM approach, we modeled the probability of being a case based on each SNP plus other potential explanatory variables, which include x1 (family history positive, yes/no); x2 (alcohol use, yes/no); x3 (tobacco use, yes/no); x4 (pickled vegetable consumption, yes/no); and x5 (age, continuous):

$$\mathrm{Prob} = 1/(1+\exp(-\mathrm{f}))\ \text{where}\ \mathrm{f(x)} = \mathrm{a} + \mathrm{b}^*\mathrm{SNP} + \mathrm{b1}^*\mathrm{x1} + \mathrm{b2}^*\mathrm{x2} + \mathrm{b3}^*\mathrm{x3} + \mathrm{b4}\ {}^*\mathrm{x4} + \mathrm{b5}^*\mathrm{x5}.$$

The three variables, tobacco use, pickled vegetable and age, were insignificant in the GLMs for nearly all SNPs and they were dropped in further analysis. Using a GLM for each SNP plus the two covariates (family history and alcohol use), we computed the *P*-value of the GLM based on the difference between the null deviance D0 and residual deviance D1 using the chi-square goodness-of-fit test. The chi-square statistic is D0–D1 with 3 degrees of freedom.

A whole genome-wide association study requires a correction for multiple comparisons. One possible adjustment for multiple testing is the use of Bonferroni adjustment, which was used in our original work. However, our more recent studies suggested that Bonferroni adjustment may not be desirable for the following reasons.

1. The reduction of type I error associated with the Bonferroni adjustment increases the type II error.
2. With the increase in SNP density such as the Affymetrix 500K SNP chip, it is unlikely to attain significant *P*-values with Bonferroni adjustment given limited sample sizes.
3. If we argue that each individual SNP has a small effect on disease, we should look for SNPs with moderate effect instead of strong effect (extremely small *P*-value) when analyzed for each SNP. More discussion can be found in **Note 5**.

Stage II in our method aims to identify interacting genetic loci, which affect disease through the joint effect of multiple genes. If we look for joint effects of genetic interactions among multiple genetic loci, each of which has a small and quantitative effect, we would naturally seek a factor that can combine these genetic loci. One solution comes from the statistical approach known as principal components analysis (PCA). If the phenotypic variation (case versus control) is primarily determined by the genetic factor, which is the result of interaction among multiple genetic loci, we would expect to find interaction of these loci affecting the likelihood of developing disease. In other words, we will find co-variation among these genetic loci in determining disease state. Such co-variation can be captured by PCA and results in clustering of samples into cases and controls. This was exactly what we found in our previous study *(13)* and in the current analysis (**Fig. 3**).

An effective way to evaluate this two-stage selection method is to use the PCA model to develop a classifier, and to assess the performance of the classifier. We used PCA to visualize sample distribution in a two-dimensional space defined by PC1 and PC2. In our analysis, case and control samples formed the two cluster structures by PC1, which can be used to construct a classifier to separate cases from controls. Our classifier is defined here as case if PC1 ≤ 0, but control if PC1 > 0. Performance of the classifier can be evaluated for accuracy as defined by (Tp+Tn)/100, sensitivity defined Tp/(Tp+Fn), and specificity defined by Tn/(Fp+Tn), where Tp and Tn are the numbers of true positives and true negatives, Fp and Fn are the numbers of false positives and false negatives.

We have discussed the need to have a moderate threshold in stage I for selecting SNPs instead of an extremely stringent *P*-value cutoff in GLM analysis to accommodate interactions among multiple genetic loci that affect the complex disease. To find a proper threshold, we ordered the SNPs with *P*-values from GLM and selected SNPs with *P*-values smaller than the cutoffs indicated in **Table 1**. We carried out PCA analysis using the number of SNPs indicated in **Table 1**. The performance of the classifier was shown in **Fig. 2**. Four examples

Table 1
Performance of the classifier based on PCA models. The definition of the terms can be found in the legend to Fig. 2. The *P*-value cutoffs used to select SNPs from GLM are provided here along with the number of SNPs selected

Number of SNPs	*P*-value cutoff	Accuracy	Sensitivity	Specificity
15	4.87E-07	0.80	0.82	0.78
31	7.72E-07	0.86	0.88	0.84
51	1.22E-06	0.86	0.84	0.88
93	1.94E-06	0.91	0.90	0.92
161	3.07E-06	0.93	0.94	0.92
278*	4.87E-06	0.94	0.94	0.94
493	7.72E-06	0.98	0.98	0.98
788	1.22E-05	0.97	0.98	0.96
1,352	1.94E-05	0.98	1.00	0.96
2,459	3.07E-05	0.97	0.98	0.96
3,393	3.90E-05	0.96	0.98	0.94
3,912	4.30E-05	0.96	0.98	0.94
4,860	4.87E-05	0.94	0.94	0.94
6,225	6.00E-05	0.86	0.88	0.84
6,604	6.50E-05	0.85	0.86	0.84
6,950	7.00E-05	0.86	0.88	0.84
7,444	7.72E-05	0.78	0.80	0.76
8,621	0.000122376	0.56	0.58	0.54
9,226	0.000193953	0.59	0.56	0.62
9,545	0.000307394	0.56	0.58	0.54
9,739	0.000487187	0.54	0.54	0.54

* Bonferroni-adjusted significance level of 0.05.

of samples projected in the 2-dimension space defined by PC1 and PC2 are shown in **Fig. 3**. When nearly all SNPs were used in PCA **(Fig. 3A)**, controls and cases intermingled with each other, indicating a homogenous population in this study. With a progressive increase in stringency in selecting SNPs (smaller *P*-value cutoff and fewer numbers of SNPs selected), we saw an increase in discrimination of case versus control in PC1 and corresponding increases in accuracy, sensitivity, and specificity **(Fig. 2** and **Table 1)**. The performance reached maximum at the *P*-value cutoff of 7.72×10^{-6}, which yielded 493 SNPs with *P*-values smaller than the cutoff **(Fig. 3B** and **Table 1)**. Further decrease in *P*-value cutoff resulted in less discrimination between case and control, and reduction in the performance. This is also true for the *P*-value

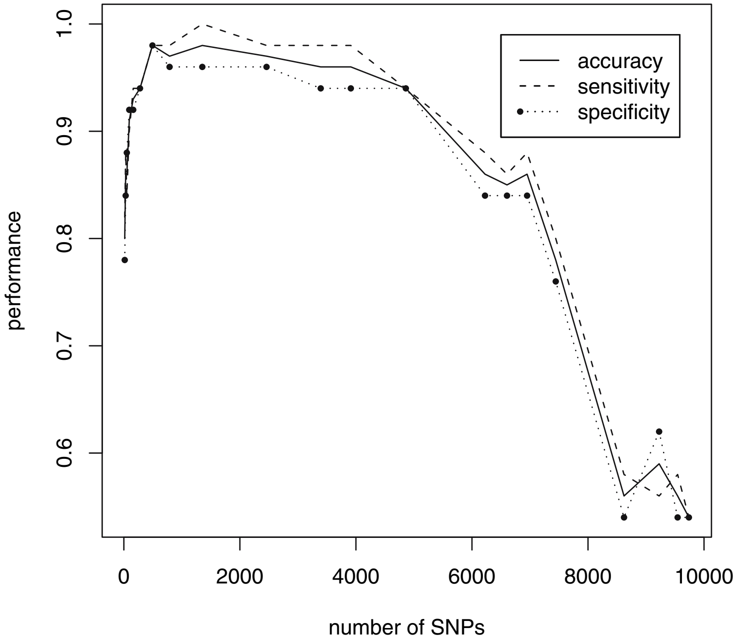


Fig. 2. Performance of the classifier based on PCA models. Three performance indices—accuracy, specificity, and sensitivity, are evaluated for classification of samples by the PC1 score. Samples are classified as control if PC1 > 0. Samples are classified as case if PC1 ≤ 0. The same data are shown in **Table 1** with more details.

cutoff that corresponded to a Bonferroni adjustment (**Fig. 3C** and **Table 1**). The *P*-value cutoff at 4.87 × 10^{-6} (p=0.05 after Bonferroni adjustment) yielded 278 genes. Further decrease in the number of SNPs generated even poorer performance (**Fig. 3D** and **Table 1**). The best performance with the *P*-value cutoff of 7.72 × 10^{-5} included 493 genes, which appear to capture most of the genes that are involved in the gene–gene interaction and contributing to disease. More discussion of the performance of the classifier can be found in **Note 5**. Furthermore, we can find the genetic risk loci from the SNPs with high loading (coefficient) in the first principal component (PC1). The absolute values of loading in PC1 are plotted in **Fig. 4** with descending order of loading on x-axis. The curve shows a steep drop in loadings and it levels off on the right side of the tail. This pattern suggests that we should look for complex disease genes in those SNPs with a large loading value (about 30 SNPs in this

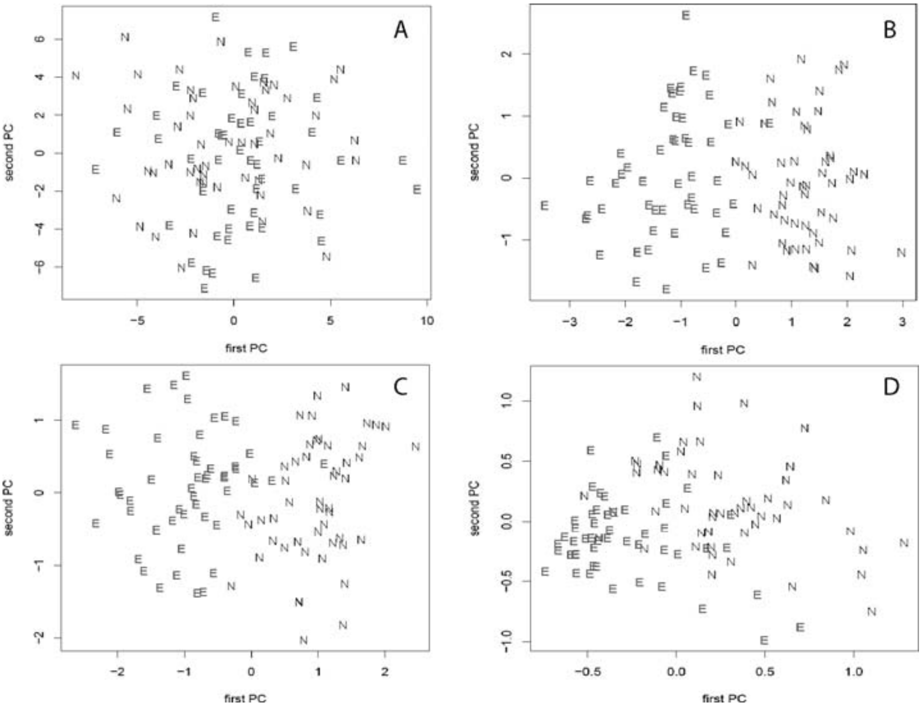


Fig. 3. Two clusters of case and control samples analyzed by PCA. **A:** 9,739 SNPs selected by *P*-value cutoff of 4.87×10^{-4} in GLM were used for PCA analysis. **B:** 493 SNPs selected by *P*-value cutoff of $7.72X10^{-6}$ in GLM were used for PCA analysis. This analysis has the best performance for accuracy, specificity, and sensitivity. **C:** 278 SNPs selected by *P*-value cutoff of $4.87X10^{-6}$ in GLM were used for PCA analysis, which corresponds to a Bonferroni adjusted significance level of 0.05. **D:** 15 SNPs selected by *P*-value cutoff of 4.87×10^{-7} in GLM were used for PCA analysis. E – cases; N – controls.

analysis) (*see* more discussion on **Note 6**). In our previous study, we used the permutation test to evaluate the performance of the PCA model ***(13)***. A more effective validation test should be done on a different set of samples. We are currently pursuing validation of selected SNPs in large numbers of external samples.

In conclusion, our two-stage sequential selection method provides an effective strategy systematically to identify susceptibility genes for complex diseases through whole genome-wide association study.

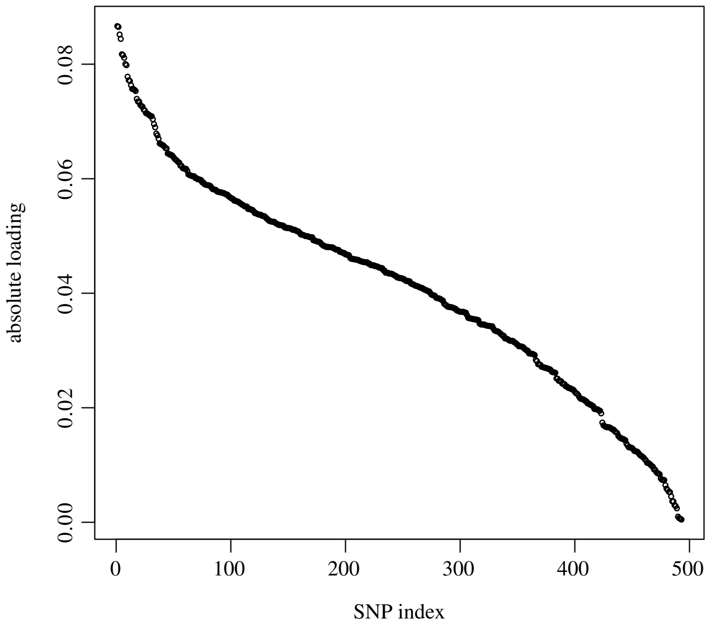


Fig. 4. Distribution of the loadings in PC1. The absolute value of the loading (coefficient) for SNP in PC1 is plotted according to the descending order. Note that decrease in the absolute value of the loadings is very steep initially and the curve flattens gradually.

4. Notes

1. Our goal in this pilot case-control study was to identify genes important for the genetic etiology of ESCC. We selected individuals with almost identical features in environmental exposures as well as demographic measures. Although small variations in those environmental effects can be controlled for in GLM analysis, it is more powerful to identify genetic factors if variations in non-genetic factors are kept minimal. If the purpose is to identify gene–environment interactions, we can introduce variations in environmental exposures in cases and controls in a balanced manner. We can identify gene–environment interactions with GLM, and require small *P*-values for the model, and moderate to small *P*-values for the coefficients of both SNP and environmental factor.
2. In addition to the standard arrays, both Affymetrix and Illumina offer custom arrays for high-throughput genotyping of selected SNPs. For genotyping of large number of samples with limited number of SNPs, it is more cost effective to use genotyping platforms such as Applied Biosystems Taqman®, SNaPshot^{TM}, and SNPlex^{TM}.

3. We found that the majority of SNPs with deviation from HWE were due to either low minor allele frequency or the fact that the SNP sequence was present in multiple genomic loci or low signal for genotype call. We evaluated the signal of the genotype call by t-test for the quantity of (PM-MM) across 20 probes for the SNP. Here PM denotes perfect match probe and MM denotes mismatch probe. The removal of those SNP with deviation from HWE reduces false positives. We used a *P*-value of 0.01 as a cutoff in the chi-square test for HWE.
4. There are many different versions of the two-stage method. The general concept is to perform genome-wide genotyping on high density SNP chips on moderate numbers of samples, usually in the range of a few hundred. Stage II focuses on interesting regions selected from stage I and involves denser genotyping with higher-density SNPs in specific regions, and on a large set of samples. The concept of a two-stage method is evolving to include selection of individual SNPs in stage I, and selection of two-locus gene interactions in stage II in whole-genome-wide association studies for complex diseases. In this paper, our two-stage approach is related to the second definition of the two-stage method. Furthermore, our method can also be applied to multi-locus gene interactions in stage II.
5. Our selection criteria in stage I rely on both a *P*-value from the GLM model and a *P*-value for the coefficient of the SNP ***(13)***. However, in the current analysis, we found that we could simply choose a moderate *P*-value from the GLM model and achieve very good performance for the PCA classifier. If the primary interest is in the genetic factor, we should pay more attention to the SNP coefficient. If the primary interest is in the gene-environment interaction, we need to pay more attention to both SNP coefficient and coefficient for the environmental factor. Case and control samples will need to include variation in the environmental factor. We recommend trying several different criteria by considering GML model and/or coefficients at different *P*-value cutoffs. The performance of selection criteria can be evaluated by the PCA classifier as described in this paper.
6. The curve shown in **Fig. 4** indicates a fairly complex dataset. The loadings show gradual decrease. The ideal situation would be a sharp drop in the absolute value of the loading for ~10–20 SNPs. Those SNPs can be further analyzed in the follow-up validation experiment

Acknowledgments

We would like to thank Drs. Alisa Goldstein and Ying Hu for discussion. This research was supported by the Intramural Research Program of the NIH, the National Cancer Institute, the Center for Cancer Research, and the Division of Cancer Epidemiology and Genetics.

References

1. Botstein, D., and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**(Suppl), 228–237.

2. Miller, M. P., and Kumar, S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10**, 2319–2328.
3. Lander, E. S. (1996) The new genomics: global views of biology. *Science* **274**, 536–539.
4. Risch, N., and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.
5. Risch, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856.
6. Campbell, D. B., Sutcliffe, J. S., Ebert, P. J., Militerni, R., Bravaccio, C., Trillo, S., et al. (2006) From the cover: a genetic variant that disrupts MET transcription is associated with autism. *Proc. Natl. Acad. Sci. U S A* **103**, 16834–16839.
7. Dewan, A., Liu, M., Hartman, S., Zhang, S., Liu, D. T., Zhao, C., et al. (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992.
8. Yang, Z., Camp, N. J., Sun, H., Tong, Z., Gibbs, D., Cameron, J., et al. (2006) A Variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. *Science* **314**, 992–993.
9. International HapMap Consortium. (2003) The international HapMap project. *Nature* **426**, 789–796.
10. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.
11. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, L. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120.
12. Li, J. Y. (1982) Epidemiology of esophageal cancer in China. *Natl. Cancer Inst. Monogr.* **62**, 113–120.
13. Hu, N., Wang, C., Hu, Y., Yang, H. H., Giffen, C., Tang, Z. Z., et al. (2005) Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. *Cancer Res.* **65**, 2542–2546.
14. Matsuzaki, H., Loi, H., Dong, S., Tsai, Y. Y., Fang, J., Law, J., et al. (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14**, 414–425.
15. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1**, 109–111.
16. Fan, J. B., Gunderson, K. L., Bibikova, M., Yeakley, J. M., Chen, J., Wickham, G. E., et al. (2006) Illumina universal bead arrays. *Methods Enzymol.* **410**, 57–73.
17. Oliphant, A., Barker, D. L., Stuelpnagel, J. R., and Chee, M. S. (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* Suppl., 56–58, *see* also pages 60–51.
18. Elston, R. C., Guo, X., and Williams, L. V. (1996) Two-stage global search designs for linkage analysis using pairs of affected relatives. *Genet. Epidemiol.* **13**, 535–558.

19. Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet*. **38**, 209–213.
20. Marchini, J., Donnelly, P., and Cardon, L. R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet*. **37**, 413–417.

3

Utilizing HapMap and Tagging SNPs

Christopher A. Haiman and Daniel O. Stram

Summary

Advancements in our understanding of variation in the human genome and rapid improvements in high-throughput genotyping technology have made it feasible to study most of the human genetic diversity that is due to common variations in relation to observable phenotypes. Over the past few years, public SNP databases have matured and empirical genome-wide SNP data, such as that generated by the International HapMap Project, have shown the utility and efficiency of selecting and testing informative markers ("tag SNPs") that exploit redundancies among nearby polymorphisms due to linkage disequilibrium (LD). In this chapter, we will demonstrate how to use the HapMap resource and the Haploview program to process and analyze genetic data from HapMap, to evaluate LD relations between SNPs, and to select tagging SNPs to be examined in disease association studies.

Key Words: single nucleotide polymorphism, linkage disequilibrium, tagging (tag) SNPs, HapMap, Haploview.

Abbreviations: D′ – dprime; HWE – Hardy Weinberg equilibrium; kb – kilobase; LD – linkage disequilibrium; MAF – minor allele frequency; populations mapped in the HapMap project include: Utah residents with Northern and Western European ancestry (CEU); Han Chinese (CHB); Japanese (JPT); Yorubans from Nigeria (YRI); Mb – megabase; SNP – single nucleotide polymorphism.

1. Introduction

Global efforts to characterize genetic sequence variation ***(1,2)*** and rapid advances in genotyping technology provide researchers with the necessary tools to decipher the contribution of inherited genetic variation to disease risk

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

in the population. In contrast to rare, highly penetrant alleles, which have a clear role in straightforward heritable forms of disease, common, less penetrant alleles have been hypothesized and more recently been validated to play an important role in many common diseases, such as diabetes ***(3)***, cancer ***(4)***, and autoimmune diseases ***(5)***, among others.

Approaches to identifying common disease alleles include testing each putative causal allele directly for association with disease risk, such as a non-synonymous single nucleotide polymorphism (SNP), in a gene positioned in a biological pathway that is closely linked with the pathogenesis of a disease. The success of this approach relies on cataloguing these alleles by re-sequencing functional domains of genes, a costly and impractical route based on the large number of genes that could be implicated in the pathogenesis of any one disease. Recently, empirical studies of human genetic variation have revealed that nearby SNPs show strong correlation (called linkage disequilibrium, or LD) that exist in long, yet highly variable, segments across the human genome ***(6)***. The coinheritance between SNP alleles enables most of the common genetic variation in a region to be captured by genotyping subsets of SNPs (termed *haplotype-tagging SNPs*, or *tag SNPs*) across a candidate gene or region of interest. These tagging SNPs are selected to predict the un-genotyped SNPs in the region (such as causal alleles), with the goal being to efficiently extract as much information about genetic variation in a region for the lowest possible cost; allowing for a more comprehensive and efficient method to test human genetic variation for association with disease risk.

This indirect genomic approach can be divided into the following steps: (1) defining the candidate region of interest, (2) empirically characterizing LD patterns across the region, (3) selecting *tagging SNPs* based on a defined set of criteria that highly predict all common variation, and (4) genotyping these markers in genetic studies to test for association with disease risk. Below we review the first three steps of this procedure. As a practical example of this approach, the genetic data from the HapMap database will be utilized as the foundation for fine-mapping chromosome 8q24 in individuals of European ancestry. This region is commonly amplified in prostate cancer tumors ***(7)***. We ***(8)*** and others ***(4)*** have also recently provided strong evidence to suggest that it harbors a susceptibility locus for prostate cancer.

2. The HapMap Project

To accelerate the identification of common disease alleles, the International HapMap Project in 2002 initiated the construction of a genome-wide SNP database of common variation ***(2)***. This publicly available resource has recently been completed and provides valuable information about LD patterns across the genome in multiple population samples. In brief, the project has genotyped over 3 million SNPs in 269 samples from four populations (90 Utah Residents (30 parent-offspring trios) with Northern and Western European Ancestry (CEU), 45 Han Chinese from Beijing, China (CHB), 44 Japanese from Tokyo, Japan (JPT), and 90 Yorubans (30 trios) from Ibadan, Nigeria (YRI)). The average spacing of the map is 1 SNP per kb, and this vast resource is currently being used globally as a template for both LD-based candidate gene and genome-wide association studies. In this chapter, we will demonstrate how to use the HapMap resource, and Haploview ***(9)***, a commonly utilized program that can process and analyze genetic data from HapMap, to evaluate LD relations between SNPs and to select tagging SNPs for disease research.

3. The HapMap Project Resource

3.1. Browsing the HapMap Project Database

1. The HapMap database is publicly available on the web at www.hapmap.org/. To access the genetic data select "Browse Project Data," listed under the heading Project Data in the left hand column of the home page. Data from the various phases of the HapMap Project are available and can be selected from the pull-down menu under Data Source. For the analysis presented below, select "HapMap Data Rel#21/phase II Jul 06, on NCBI assembly, dbSNP b125" (*see* **Note 1**).
2. To search the HapMap database, gene name (common name or NCBI RefSeq accession number), dbSNP rs#, chromosome name or band, or genetic coordinates of the candidate region of interest must be provided under Landmark or Region. In this example, we will limit our investigation to a 500-kb area located within a 3.8-Mb region at chromosome 8q24, from 125.68 to 129.48 Mb in build 35 of the human genome sequence (13.9 cM), a region that has recently being implicated in contributing to prostate cancer susceptibility. To access data for this segment type "chr8:127325000..127824999" under Landmark or Region. Select the tab "Search" to execute the request (*see* **Note 2**).
3. Under Overview, summary information for chromosome 8 is provided, including the cytogenetic chromosome bands, the density of genes in the NCBI Entrez Gene database as well as the number of SNPs in dbSNP that were genotyped by the HapMap Project per 500 kb window. The red rectangle designates the region of interest which is magnified below under the heading Details. Here, the SNPs genotyped by the HapMap Project in 20 kb windows are shown, as are LocusLink genes and mRNA sequences (e.g., NM_ 174911), which indicate annotated genes and putative functional regions featured in NCBI's RefSeq database.
4. On the right of the page, the Scroll/Zoom arrows, buttons and drop-down menu can be used to magnify or minimize the view and reposition the display window along the chromosome. Select "Show 20 kbp" from the drop down menu. More

detail is now provided for each SNP genotyped in this 20 kb window, including the SNP rs#, and SNP frequency in each of the HapMap populations (CEU, CHB, JPT and YRI), illustrated as colored pie charts with different colors (blue or red) representing the two different alleles for each SNP. Clicking on a SNP provides detailed information pertaining to counts and frequencies of genotypes and alleles as well as assay information. For a broader view, select "Show 100 kbp" from the drop down menu; each genotyped SNP is now shown as a triangle (*see* **Note 3**). To reposition the display to the left, click on the double arrow "«". This will re-center the graphical viewer to a 100-kb region spanning 127,425,000 to 127,524,999 as illustrated in **Fig. 1** (Top).

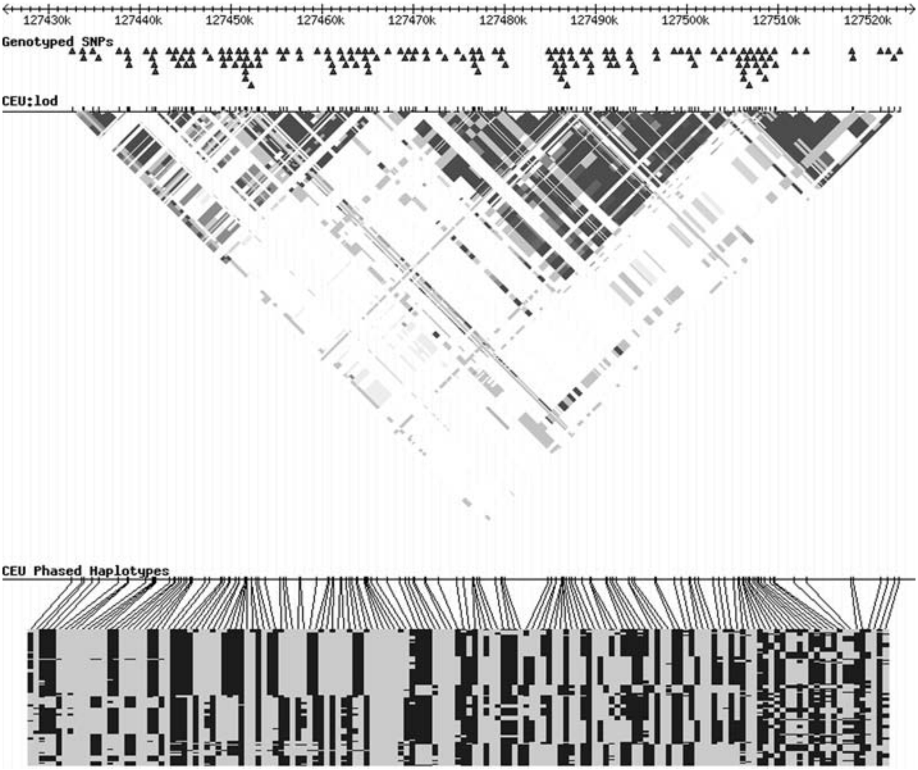


Fig. 1. Genome browser displays in HapMap. **Top:** The SNPs genotyped by the HapMap Project spanning a 100 kb region on chromosome 8q24 (127,425,000 - 127,524,999). **(Middle)** A linkage disequilibrium plot for these SNPs. **Bottom:** The phased haplotypes estimated for the CEU population in this region.

3.2. Defining Linkage Disequilibrium Patterns in HapMap

1. The categories listed under Tracks at the bottom of the page enable additional features to be displayed in the Details section. Check the box “plugin:LD plot” in the Analysis section to view a graphic representation of LD patterns between SNPs in the region. Refresh the page by clicking any of the “Update Image” buttons. A linkage disequilibrium (LD) plot of the association between SNPs (represented as boxes) is now shown below, with the color intensity of each box depicting the strength of these relationships; here the white boxes indicate recombination has occurred and that there is little or no LD between SNP pairs. This plot is shown for the CEU population, which is the default setting. Here we see multiple regions of LD across this 100-kb segment, with the largest segment or “block” of LD spanning ~35 kb (127,470k–127,505k). A linkage disequilibrium plot for the CEU population is shown in **Fig. 1** (Middle).
2. To annotate the LD plot select “Annotate LD Plot” listed as one of the options in the pull down menu under Reports & Analysis. Then click “Configure.” Here, one has the option to adjust the display settings of the LD plot, based on the size of the region examined, the number of SNPs genotyped, and whether the box size in the plot is displayed in a uniform size or proportional to the genomic distance between SNPs. Other parameters include which LD measure to apply (D′, r^2 or LOD score) (LD Properties) (*see* **Note 4**), the range of LD values to define strong LD, and which population samples to display (Populations). The default color scheme is based on the combination of D′ and LOD score and is the same as the standard color scheme provided in Haploview (discussed below). As an example, select “dprime” in the LD Properties pull-down menu and specify “greater than 0.8 and less than 1.0” in the adjacent pull down menu. Turn all populations “on” and set all orientations to “normal.” Select “Configure” to process (*see* **Note 2**). LD patterns for CEU, CHB, JPT and YRI populations are now shown separately.

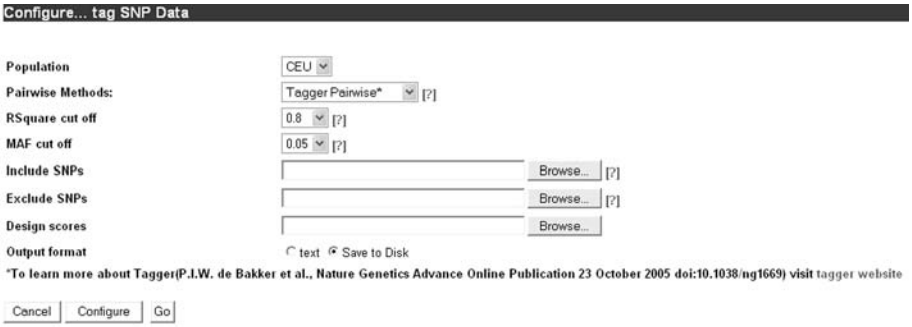
3.3. Haplotype Patterns in HapMap

Phased haplotypes can be displayed by checking the box “plugin:Phased Haplotype Display” in the Analysis section. Refresh the page by clicking “Update Image.” Phased haplotypes are estimated in each population using a maximum likelihood algorithm (JPT and CHB populations are combined) (*see* **Note 5**). Haplotypes shown are represented by two colors (blue or yellow) for each SNP allele. Haplotypes are shown for each subject (horizontally), with regions of high LD represented by long stretches of the same color. Phased haplotypes for CEUs across this 100 kb region are shown as an example in **Fig. 1** (bottom).

3.4. Selecting Tag SNPs in a 50-kb Region at 8q24

HapMap uses algorithms in the Tagger program *(10)* to select tag SNPs. More information about this program is available at the Tagger website (www.broad.mit.edu/mpg/tagger/).

1. To select tagging SNPs across the 500 kb region of interest first reposition the display by typing “chr8:127325000..127824999” under Landmark or Region and click ‘Search.” The graphical LD and haplotype displays are not needed for tag SNP selection so unselect all of the boxes listed under “Analysis” in the Tracks section below. The HapMap Project allows users to customize the settings used to select tagging SNPs. This can be done by selecting “Download tag SNP data” in the Reports & Analysis section and then “Configure.” Here one can set preferences (**Fig. 2**) such as the population group (*see* **Note 6**), the tag SNP selection algorithm to be implemented (*see* **Note 7**), the pairwise correlation threshold (r^2 value) for predicting SNPs (*see* **Note 8**), and the minor allele frequency (MAF) cutoff of SNPs to predict (*see* **Note 9**). For the present analysis select, Population: CEU, Pairwise Methods: Tagger Pairwise, Rsquare cutoff: 0.8, and MAF cut off: 0.05. SNPs can also be included or excluded preferentially as tagging tags by uploading additional .txt files (*see* **Note 10**). To generate a report that summarizes the tag SNP selection procedure choose “Save to disk” and then click the “Go” button. The file will be saved to your local computer as a text (.txt) file. Open the file just saved in Microsoft Excel (recommended). Three pieces of information are provided in the report. The first section is titled, “HapMap tag SNPs:127 tag SNPs picked out for


Configure... tag SNP Data

Population: CEU

Pairwise Methods: Tagger Pairwise* [?]

RSquare cut off: 0.8 [?]

MAF cut off: 0.05 [?]

Include SNPs: [?]

Exclude SNPs: [?]

Design scores

Output format: ○ text ● Save to Disk

*To learn more about Tagger(P.I.W. de Bakker et al., Nature Genetics Advance Online Publication 23 October 2005 doi:10.1038/ng1669) visit tagger website

Fig. 2. Picking tagging SNPs in HapMap. The tag SNP configuration page is an interactive page that allows users to select tagging SNPs, using the Tagger program, based on a number of user defined criteria. These custom settings include the population of interest, the tag SNP selection algorithm to be implemented, the pairwise correlation threshold (r^2value) for predicting SNPs and the minor allele frequency (MAF) cutoff of SNPs to predict, among others.

population CEU chr8:127325000..127824999 using the algorithm-Tagger-pairwise Tagging." This report lists the 127 tag SNPs selected to predict all genotyped SNPs (MAF $\geq$ 0.05) in HapMap for the CEU population across this region as well as their genome position and MAF frequency (*see* **Note 11**). In the second section, the number of SNPs predicted by the tagging SNPs is indicated (n = 456) as well as the mean r^2 value for predicting all SNPs (0.956). What follows is the list of these 456 common SNPs ("Marker"), the best tag SNP predictor of each SNP ("Best Test") and the corresponding r^2 value. The last section reports each tag SNP ("Test") and the SNPs that are predicted by the tag with an $r^2 \geq 0.8$ ("Alleles Captured").

2. Using the information provided in this file one can calculate the density of common SNPs (MAF $\geq$ 0.05) genotyped by the HapMap Project across any region for each population sample. For whites, 1 common SNP was genotyped every 1,096 base pairs on average (456 SNPs across 500 kb).

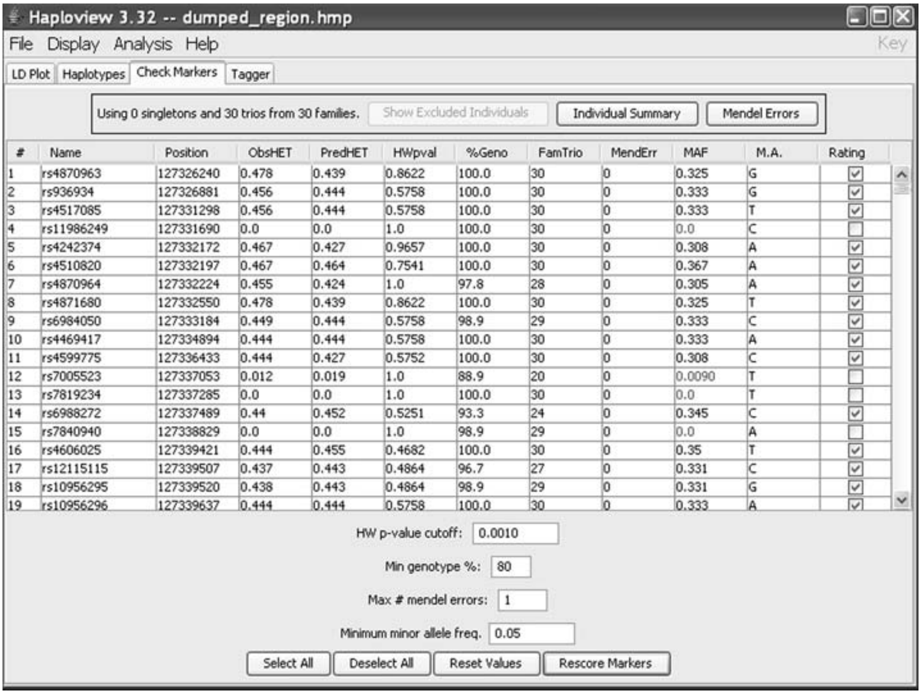
3.5. Evaluating HapMap Data Using Haploview

1. Haploview is a commonly used program for processing and analyzing genetic data and has been adapted to handle data from the HapMap Project *(9)*. This program is publicly available at www.broad.mit.edu/mpg/haploview/. To download the program, go to the website and click on "download" in the section Haploview version 3.32 (a later version may also be available). Haploview uses the Java Runtime Environment (JRE). If this has not been previously installed, click on the link www.java.com/ and follow the instructions to download the newest version of the software. Next, download the Haploview program for your operating system and run the Haploview Executable Jar File (the .exe file) to install the program.
2. To use Haploview, first download the genotype data from the HapMap database. To do this, select "Browse Project Data," listed under Project Data at the HapMap home page. Enter "chr8:127325000..127824999" under Landmark or Region. Then choose "Download SNP genotype data" in the Reports & Analysis menu and click "Configure." Here, one can select the genetic data to be downloaded for a specific population and the strand on which the SNP alleles will be presented (rs, does not specify strand; fwd, forward strand relative to latest NCBI genome build; rev, reverse strand in build). Select the CEU population, rs strand and "Save to disk." Then click "Go." The file will be saved as a .hmp file.
3. Open the Haploview program by clicking on the Haploview icon (located in the directory in which the program was saved). Once opened, there will be three choices for loading genetic data into Haploview. Select "Load HapMap Data" (the tutorial provided in the Haploview home page presents the other two file format options). In the "Genotype data" panel, one can Browse for or type the path where the HapMap data (.hmp file) was saved. Below, enter "500" in first box and "50" in the second box; these are the default settings for the distance over which pairwise relationships

between SNPs will be evaluated, and the percentage of missing data allowed per subject, respectively. Click "OK" to proceed.

3.6. SNP Descriptions and Quality Control Checks in Haploview

Description information about each SNP is presented in the tab "Check Markers" (**Fig. 3**). Here, SNPs are numbered based on their position provided in the uploaded file (as shown in the first column). This SNP numbering scheme is used to identify the SNP in all subsequent Haploview routines. This display also shows the rs#, genome position, p-value for Hardy Weinberg equilibrium (HWE), genotyping percentage, and MAF for each SNP, as well as other descriptive information. In the last column, the unchecked boxes indicate those SNPs that will not be further analyzed based on quality control filtering,



Using 0 singletons and 30 trios from 30 families.

#	Name	Position	ObsHET	PredHET	HWpval	%Geno	FamTrio	MendErr	MAF	M.A.	Rating
1	rs4870963	127326240	0.478	0.439	0.8622	100.0	30	0	0.325	G	☑
2	rs936934	127326881	0.456	0.444	0.5758	100.0	30	0	0.333	G	☑
3	rs4517085	127331298	0.456	0.444	0.5758	100.0	30	0	0.333	T	☑
4	rs11986249	127331690	0.0	0.0	1.0	100.0	30	0	0.0	C	□
5	rs4242374	127332172	0.467	0.427	0.9657	100.0	30	0	0.308	A	☑
6	rs4510820	127332197	0.467	0.464	0.7541	100.0	30	0	0.367	A	☑
7	rs4870964	127332224	0.455	0.424	1.0	97.8	28	0	0.305	A	☑
8	rs4871680	127332550	0.478	0.439	0.8622	100.0	30	0	0.325	T	☑
9	rs6984050	127333184	0.449	0.444	0.5758	98.9	29	0	0.333	C	☑
10	rs4469417	127334894	0.444	0.444	0.5758	100.0	30	0	0.333	A	☑
11	rs4599775	127336433	0.444	0.427	0.5752	100.0	30	0	0.308	C	☑
12	rs7005523	127337053	0.012	0.019	1.0	88.9	20	0	0.0090	T	□
13	rs7819234	127337285	0.0	0.0	1.0	100.0	30	0	0.0	T	□
14	rs6988272	127337489	0.44	0.452	0.5251	93.3	24	0	0.345	C	☑
15	rs7840940	127338829	0.0	0.0	1.0	98.9	29	0	0.0	A	□
16	rs4606025	127339421	0.444	0.455	0.4682	100.0	30	0	0.35	T	☑
17	rs12115115	127339507	0.437	0.443	0.4864	96.7	27	0	0.331	C	☑
18	rs10956295	127339520	0.438	0.443	0.4864	98.9	29	0	0.331	G	☑
19	rs10956296	127339637	0.444	0.444	0.5758	100.0	30	0	0.333	A	☑

HW p-value cutoff: 0.0010

Min genotype %: 80

Max # mendel errors: 1

Minimum minor allele freq. 0.05

Fig. 3. Check markers in Haploview. The Check Markers display in Haploview allows users to filter out SNPs based on specified quality control criteria which include threshold for deviation from Hardy Weinberg equilibrium (HWE), minimum genotyping success rate, number of Mendelian inheritance errors, and minor allele frequency.

e.g., HWE p-value, number of Mendelian inheritance errors or missing data, or minor allele frequency. These criteria can be specified below. Enter 0.001 for "HW *p*-value cutoff"; 80 for "Min genotype %"; 1 for "Max # mendel errors"; and 0.05 for "Minimum minor allele freq." Then click "Rescore Markers." The data loaded into Haploview came directly from HapMap, so the first three filters have already been applied to the data. Subject exclusion criteria differ slightly between HapMap and Haploview which could affect SNP filtering at this step (*see* **Note 12** before proceeding). Information in the table presented can be saved by selecting "Export current tab to text" under File on the toolbar and provide a file name. To view, open the resulting file in Microsoft Excel.

3.7. Linkage Disequilibrium Plots in Haploview

LD relations between the 456 common SNPs (MAF ≥ 0.05) (*see* **Note 12**) across this 500 kb region are displayed graphically in the first tab, "LD plot" (**Fig. 4**). At the top of the page are the rs#'s for each SNP, SNP #, and their positions relative to one another; only those SNPs checked in the "Check Markers" table are presented. The map in the lower left corner provides an overview of the LD pattern for the region. Clicking different spots on this map allows one to navigate quickly to different areas of the plot, with the region enclosed by the black rectangle being magnified in the presentation above. In the enlarged map, the color of each box signifies the strength of the relationship between SNP alleles. Here, the LD color display is based on the D′/LOD score ratio ("Standard (D′/LOD)") which is the default setting in Haploview. Alternative LD presentation/color schemes can be selected under "LD color scheme" located in Display on the toolbar (*see* **Note 4**) and a key for each color scheme can be found in the upper right corner of the screen. Right click in any box in the LD plot to present a detailed summary of LD relations between SNPs, including the distance between SNPs, D′ value, D′ confidence bounds, LOD score, and r^2 value. Regions of strong LD, commonly referred to as "LD Blocks" ***(6)***, can be defined in Haploview using a number of algorithms, including the confidence interval of D′ (which is the default setting ***(6)***), the 4-gamete rule, the solid spline of LD, or based on user defined criteria. To modify block definitions, select "Define Blocks" under the Analysis tab on the toolbar. Blocks are shown on the LD plot as black triangles and denote regions where there is little evidence of recombination. Block numbers as shown above each block moving left to right across the region, with block size being dependent on the number of SNPs typed and extent of LD. Blocks can also be edited by hand by clicking on a SNP number (located below the rs#), and while holding down, dragging the cursor to the right or left; this

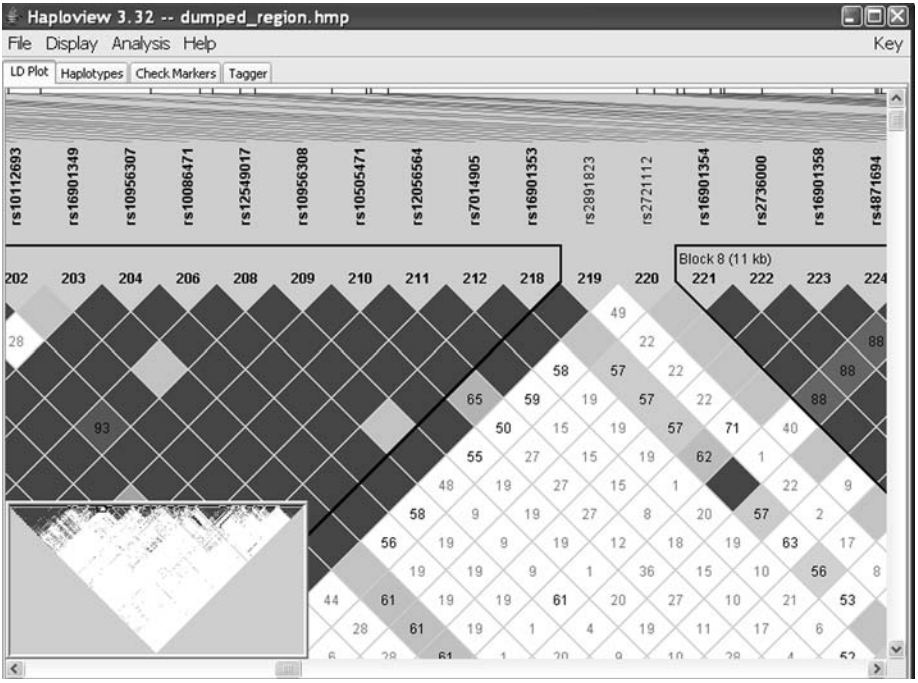
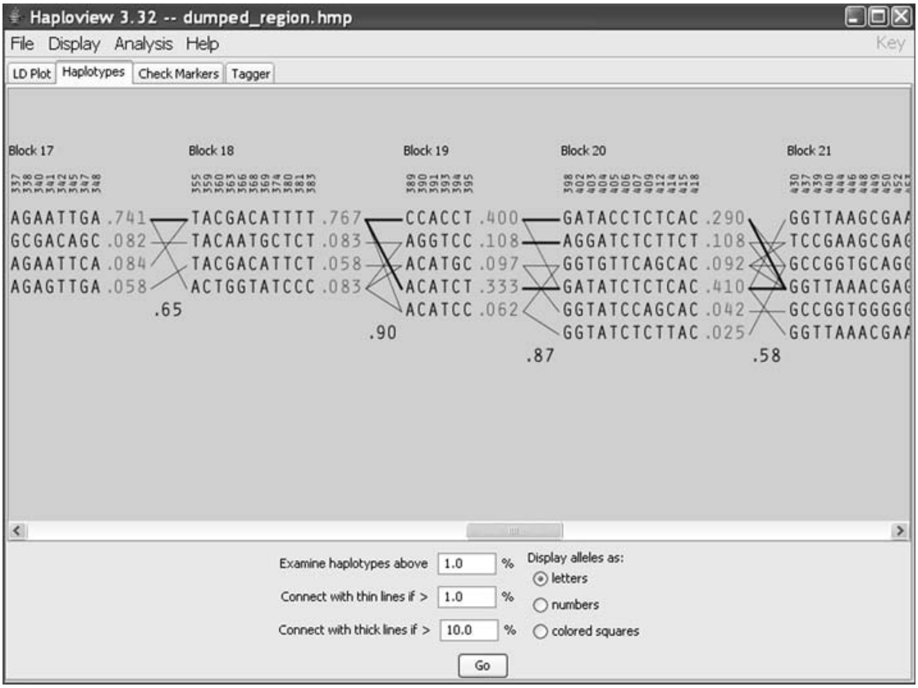


Fig. 4. LD display in Haploview. The LD plot display in Haploview shows a high-resolution illustration of the LD relations between SNPs selected in the Check Markers tab. The shade of each square indicates the strength of the LD relationship between pairs of SNPs. A broad overview of the LD patterns is shown in the left-hand corner of the screen.

redefines the boundaries of the block. In this 500 kb region at 8q24, 35 LD blocks are noted.

3.8. Viewing Haplotypes in Haploview

Click on the "Haplotypes" tab to view the haplotypes within defined LD blocks (**Fig. 5**). Haplotypes are estimated from the genotype data using an estimation-maximization algorithm *(11)*. This display shows the haplotypes in each LD block and their estimated frequency. SNP # is provided above the haplotypes and A, C, G and T indicates the nucleotides for each SNP allele. Again, only those SNPs checked in the "Check Markers" table are presented. The width of the black lines between blocks represents the strength of the correlation between haplotypes while the number located between adjacent



Examine haplotypes above 1.0 %
Connect with thin lines if > 1.0 %
Connect with thick lines if > 10.0 %

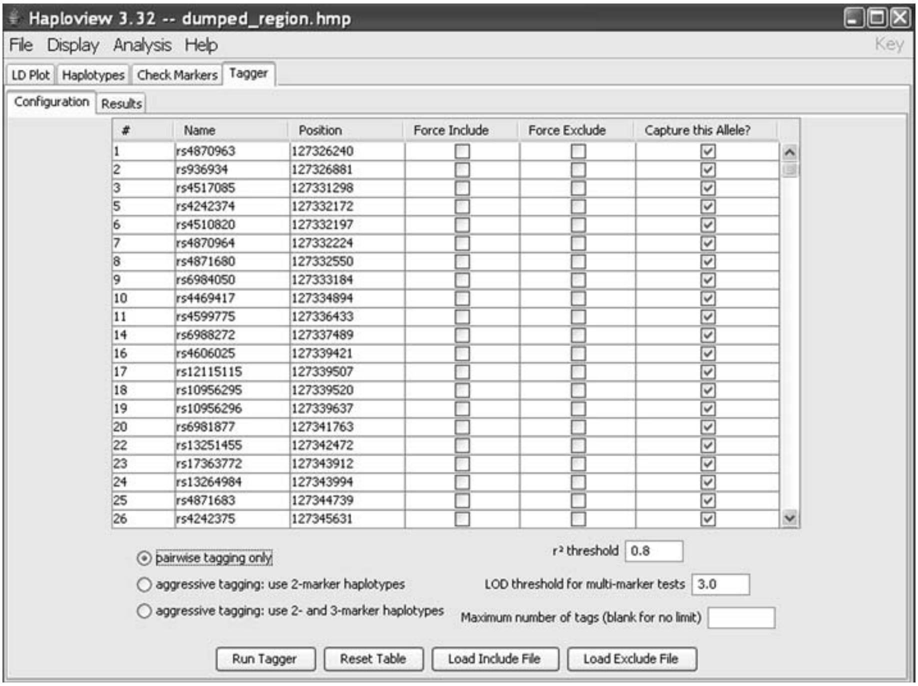
Display alleles as:
● letters
○ numbers
○ colored squares

Fig. 5. Viewing haplotypes in Haploview. The haplotype display in Haploview shows each haplotype within defined LD Blocks, their estimated population frequency, and relations with haplotypes in adjacent Blocks.

blocks is a multiallelic D′ value *(12)* which is a measure of the level of recombination between blocks. At the bottom of the screen, one can modify the display to show only those haplotypes above a specified frequency, and, the presentation of the inter-block haplotype relations and the SNP alleles (as a letter, number or color).

3.9. Selecting Tagging SNPs in Haploview

1. Haploview, like HapMap, also utilizes the Tagger program for selecting tagging SNPs. Select the tab “Tagger,” located at the top right of the screen (the fourth tab from the right). This will present two additional tabs titles “Configuration” and “Results.” In the “Configuration” display (**Fig. 6**) all SNPs checked in the “Check Markers” tab will be listed. The table shows SNP#, rs#, and position of each SNP as well as three checkbox options for each SNP: “Force Include” allows SNPs to be forced in as a tag SNP, “Force Exclude” prohibits a SNP from being selected as a tag SNP, and “Capture this Allele” designates which SNPs are to be predicted (this


Haploview 3.32 -- dumped_region.hmp
File Display Analysis Help
Key
LD Plot Haplotypes Check Markers Tagger
Configuration Results
Run Tagger
Reset Table
Load Include File
Load Exclude File

#	Name	Position	Force Include	Force Exclude	Capture this Allele?
1	rs4870963	127326240	□	□	☑
2	rs936934	127326881	□	□	☑
3	rs4517085	127331298	□	□	☑
5	rs4242374	127332172	□	□	☑
6	rs4510820	127332197	□	□	☑
7	rs4870964	127332224	□	□	☑
8	rs4871680	127332550	□	□	☑
9	rs6984050	127333184	□	□	☑
10	rs4469417	127334894	□	□	☑
11	rs4599775	127336433	□	□	☑
14	rs6988272	127337489	□	□	☑
16	rs4606025	127339421	□	□	☑
17	rs12115115	127339507	□	□	☑
18	rs10956295	127339520	□	□	☑
19	rs10956296	127339637	□	□	☑
20	rs6981877	127341763	□	□	☑
22	rs13251455	127342472	□	□	☑
23	rs17363772	127343912	□	□	☑
24	rs13264984	127343994	□	□	☑
25	rs4871683	127344739	□	□	☑
26	rs4242375	127345631	□	□	☑

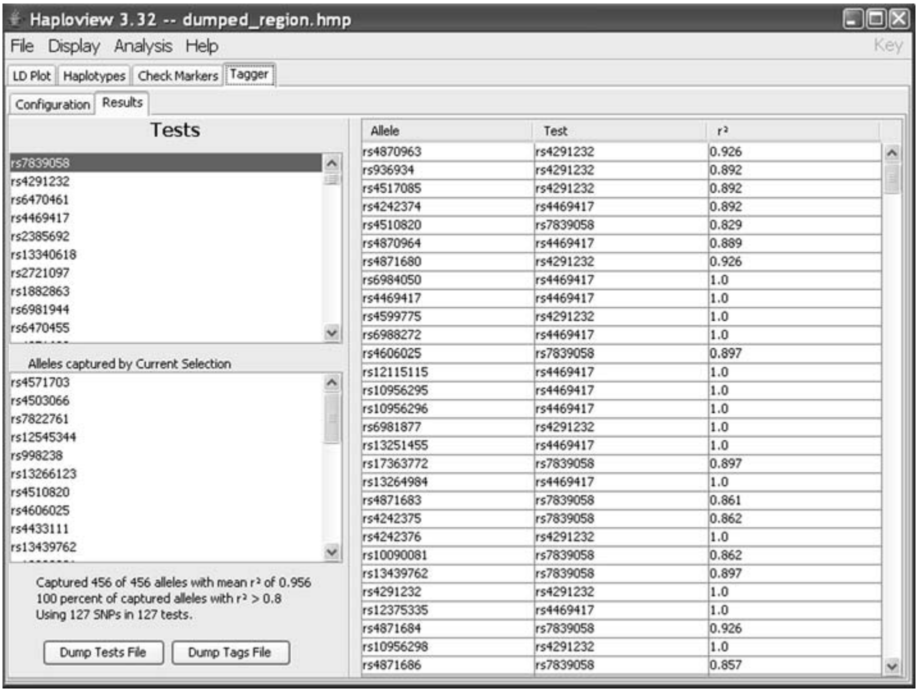
● pairwise tagging only
○ aggressive tagging: use 2-marker haplotypes
○ aggressive tagging: use 2- and 3-marker haplotypes

r^2 threshold 0.8
LOD threshold for multi-marker tests 3.0
Maximum number of tags (blank for no limit)

Fig. 6. The tagger configuration panel in Haploview. This panel shows the SNPs that will be used in the tag SNP selection process. Here, the user can select SNPs to be forced in as a tag SNP, force out SNPs from being selected as a tag, and designate which of the SNPs are to be predicted. Tagging options including pairwise and aggressive tagging strategies can be selected at the bottom of the screen.

feature is not provided in HapMap). At the bottom of the screen, tagging options include pairwise and aggressive tagging (*see* **Note 7**) and the r^2 and LOD score (used for aggressive tagging) thresholds for selecting tag SNPs. Choose "pairwise tagging only" and set "r^2threshold" to 0.8 (*see* **Note 8**). Click "Run Tagger" to run the program.

2. The "Results" panel will be displayed (**Fig 7**). On the right is a table that includes all 456 SNPs listed by SNP#. The next columns show each SNP (or test) that is the best predictor of each SNP and their r^2 value, respectively. If a SNP was left unchecked in the "Capture this Allele" column on the "Configuration" panel, then it will appear grayed out, while SNPs will appear in red if they could not be successfully tagged. There should be no SNPs highlighted in gray or red in the current display. On the top left, the "Tests" section shows the tag SNPs selected by the program. Clicking on any one of these SNPs provides a list in the lower half



Tests

rs7839058
rs4291232
rs6470461
rs4469417
rs2385692
rs13340618
rs2721097
rs1882863
rs6981944
rs6470455

Alleles captured by Current Selection

rs4571703
rs4503066
rs7822761
rs12545344
rs998238
rs13266123
rs4510820
rs4606025
rs4433111
rs13439762

Captured 456 of 456 alleles with mean r² of 0.956
100 percent of captured alleles with r² > 0.8
Using 127 SNPs in 127 tests.

Allele	Test	r²
rs4870963	rs4291232	0.926
rs936934	rs4291232	0.892
rs4517085	rs4291232	0.892
rs4242374	rs4469417	0.892
rs4510820	rs7839058	0.829
rs4870964	rs4469417	0.889
rs4871680	rs4291232	0.926
rs6984050	rs4469417	1.0
rs4469417	rs4469417	1.0
rs4599775	rs4291232	1.0
rs6988272	rs4469417	1.0
rs4606025	rs7839058	0.897
rs12115115	rs4469417	1.0
rs10956295	rs4469417	1.0
rs10956296	rs4469417	1.0
rs6981877	rs4291232	1.0
rs13251455	rs4469417	1.0
rs17363772	rs7839058	0.897
rs13264984	rs4469417	1.0
rs4871683	rs7839058	0.861
rs4242375	rs7839058	0.862
rs4242376	rs4291232	1.0
rs10090081	rs7839058	0.862
rs13439762	rs7839058	0.897
rs4291232	rs4291232	1.0
rs12375335	rs4469417	1.0
rs4871684	rs7839058	0.926
rs10956298	rs4291232	1.0
rs4871686	rs7839058	0.857

Fig. 7. The tagger results panel in Haploview. On the right is a table of the SNPs listed in the Tagger Configuration panel, the best test, i.e, predictor, of each SNP and their r^2 value, respectively. On the top left, the “Tests” section shows the tag SNPs selected by Tagger. Selecting one of these SNPs will show a list in the lower half of the panel of the SNPs captured by that SNP. At the bottom left of the screen is a summary of the tagging results.

of the panel of the SNPs captured by that SNP. At the bottom left of this panel a summary of the tagging SNP picking procedure is provided. Here the number of alleles captured by the tags was 456 and the mean r^2 = 0.956. The fraction of the alleles captured was 100% based on an r^2 threshold of ≥ 0.8. In the last line, the number of tag SNPs are indicated (n = 127) as are the number of association tests required to study common variation in HapMap predicted by this set of tags.

3. At the bottom of the page, the “Dump Tests File” button exports a list of all tests selected by Tagger for association testing while the “Dump Tags File” button exports a list of all tag SNPs selected by Tagger. For pairwise tagging these files are identical. Similar summary information as that provided in HapMap can be retrieved by selecting “Export Tab to Text” in the File menu. Open the saved file in Microsoft Excel. The first section provides a list of the 456 common SNPs

("Allele") and the tag SNP selected to predict each SNP ("Best Test"). Below in this report is a list of each tag SNP ("Test") and the SNPs that are predicted by the tag with an $r^2 \geq 0.8$ ("Alleles Captured") (*see* **Note 11**).

4. Conclusions

We have shown how to use two important tools for analysis of LD structure and tag SNP selection, and given as an example tag SNP selection for the region of chromosome 8 (8q24) suspected to harbor one or more prostate cancer-causing genetic variants. The goals of further genetic analyses that follow the tag SNP selection is to (1) localize and (2) identify the specific variants that affect risk of disease as well as (3) define their biological modes of action. The immediate next step then is to perform case-control studies (here of prostate cancer) in which the selected tag SNPs will be genotyped in cases and controls that are well matched by ethnicity, age, and sometimes other risk factors. Analyses range from simple allele counting (the frequency of each SNP in cases versus controls) through to more elaborate SNP-based logistic regression analysis (fitting various penetrance models, multi-SNP models, etc.) and haplotype-based logistic regression ***(13)***. These "association based" genetic studies generally will reduce the uncertainty about the location of a disease-causing variant from approximately 1 to 10 Mb (with linkage studies), to regions with an extent of 5–20 kb or sometimes even less. Once a specific tag SNP or tag SNP haplotype with the largest "effect" on risk is identified the task of identifying and characterizing the specific variants, e.g. through sequencing of DNA from cases carrying the specific tag SNPs, and protein and RNA analysis, etc., follows. These and other issues are treated in other contributions (*see* **Chapters 1, 11, 13**).

5. Notes

1. The SNP data in HapMap was produced in multiple genotyping laboratories using different analytic platforms and thus, quality control was of paramount concern. Strict quality control filters were applied to the data prior to making it publicly available. Each SNP must have achieved the following standards: a completion rate $\geq$ 80%, a HWE P $\geq$ 0.001, $\leq$ 1 Mendelian error (for CEU and YRI trios) and $\leq$ 1 discrepancy across 5 duplicates ***(2)***. Additional information about genotyping and quality control protocols can be found at: www.hapmap.org/downloads/data-handling_protocols.html.
2. The information entered under Landmark or Region and when configuring the settings under Reports & Analysis is stored in a browser cookie. Because this information is stored these configurations do not need to be re-entered each time.

3. The color of each triangle can be customized based on SNP properties (coding SNPs: non-synonymous versus synonymous; non-coding SNPs: intronic). To customize, select “Highlight SNP Properties” listed as one of the options in the pull down menu under Reports & Analysis and then “Configure”.
4. Many different statistics have been developed to assess LD between SNPs ***(14)*** and some of the more common measures include D′, r^2, and LOD score. In both HapMap and Haploview, LD can be shown graphically in many ways. The default settings are the D′/LOD score ratio, where the different colors indicate the strength of the relationship between SNPs.
5. Additional information about the measures employed by HapMap can be found under the *help link*, www.hapmap.org/gbrowse_help.html#LD. Additional information about the haplotype estimation procedures employed by HapMap can be found at *help link*, www.hapmap.org/gbrowse_help.html#phased_haps.
6. Genetic studies often include multiple racial and ethnic population samples. In such studies, iterative tag SNP selection procedures are recommended for selecting an optimal set of tags to be tested across populations. For example, in a study consisting of both African Americans and whites, tags should first be selected in one population (such as the CEU samples). The selected tags could then be “forced in” as tags in the second population (YRI samples), followed by the selection of additional population-specific tags required for the second group. This procedure eliminates genotyping redundant SNPs that may be selected if tags are chosen separately for each population.
7. Different methods and a variety of software have been developed for selecting tag SNPs (reviewed in ***(15)***). These basically fall into pairwise tagging (Tagger, as we have discussed, and also LDSelect, ***(16)***), multi-SNP tagging, and haplotype-based tagging. In multi-SNP tagging as implemented by Tagger and others. ***(17)*** two or more SNPs are used to construct predictors of other SNPs; this is termed by Tagger as “aggressive” tagging. In haplotype-based tagging (as in the program tagSNPs) ***(13)*** the prediction of common haplotypes is formalized by defining a special R^2 statistic. This method is “block based”; it predicts the common haplotypes seen within blocks of high LD and ignores SNPs falling into regions of low LD. Generally either multi-SNP or haplotype based tagging is more efficient than pairwise tagging in regions of high LD, although these methods introduce some additional complications to the case-control analyses.
8. An r^2 value of 0.8 is the pairwise threshold used in most studies. However, the value applied depends on the size of the association study in which the tags will be tested and available genotyping funds. The statistical power to detect an association with an unknown causal allele depends on the correlation between the selected tag SNP and the unknown variant, and the size of the study. Applying a lower r^2value must be compensated for by increasing the sample size by $1/r^2$ (considered the sample size inflation factor). Thus, for smaller studies, it is recommended to use a higher r^2 cutoff for selecting tagging SNPs (to increase the effective

sample size). However, this may result in selecting more tag SNPs to genotype in the association study which increases the overall costs of the study. Thus, there is a tradeoff between information/power and cost that must be considered with planning LD-based studies.

9. The MAF should be set no lower than 0.05 because the HapMap Project did not target SNPs with allele frequencies lower than 0.05. The HapMap resource was established to address the hypothesis that common alleles that are shared across populations underlie many common phenotypes including susceptibility to disease. Rare alleles are also likely to be important, and other efforts will be needed to find and catalogue them for study.
10. HapMap allows for the inclusion, exclusion or preferential selection of tagging SNPs. In SNP tagging studies it is customary to "force in" SNPs that should be tested directly. These SNPs may include non-synonymous coding SNPs (which were prioritized for genotyping in HapMap) or SNPs in functional regions such as in gene promoters or regions that are highly conserved across species. A SNP may also be deselected ("forced out") or preferentially selected as a tag based on other criteria, such as the potential to be genotyped successfully. For example, some genotyping platforms, i.e, Illumina, provide a design score for SNP genotyping assays which can be used to prioritize SNPs to select during the tag selection procedure. To use these additional features, an auxillary text file (.txt) must be created that lists the SNP IDs (rs#) which can be uploaded on the tag SNP configuration page.
11. The Tagger programs that appear both in Haploview and built into HapMap may give different resulting sets of tags, even when using the same settings. This is because the search algorithm used to pick the SNPs actually has some stochastic (random) aspects to it, so that if there are two SNPs that would perform equally well as tags, the version of Tagger in HapMap might pick one, and in Haploview the other. These alternatives are generally speaking equal in terms of the number of tag SNPs and the efficiency of the tags picked. In fact, the version of Tagger in Haploview can even give different results when run more than once on the same data, again because of the stochastic algorithm used, whereas in HapMap the same random choices are always used by Tagger when running the same data.
12. Unlike in HapMap, in Haploview trios with Mendelian inheritance errors are excluded from further analysis. This may result in differences in the MAFs and genotyping success rates calculated for individual SNPs and thus, the number of SNPs that pass the filtering criteria in each program. In this example, of the 671 SNPs dumped from HapMap and loaded into Haploview, 455 SNPs are estimated to have a MAF $\geq$ 0.05 and a genotyping success rate $\geq$ 80%, corresponding to 455 boxes that will be checked. This is 1 less than the number of SNPs that passed the filtering criteria in HapMap (n = 456); SNP rs11985629 was excluded in Haploview because of the issue described above. A Mendelian error resulted in a drop in the genotyping call rate to 76.7% (it was 80% in HapMap). To maintain

consistency in the number of SNPs used in the analyses that follow (n = 456), check the box in the "Rating" column to include this SNP.

Acknowledgments

We would like to thank the investigators of the HapMap Project and the developers of the HapMap website for allowing us to use screen shots of their website. We would also like to acknowledge Jeffery Barrett, Mark Daly, and others in the Daly Lab who developed and maintain the Haploview program.

References

1. International Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
2. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.
3. Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323.
4. Amundadottir, L. T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A., Agnarsson, B. A., et al. (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658.
5. Ueda, H., Howson, J. M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., et al. (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511.
6. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.
7. Sato, K., Qian, J., Slezak, J. M., Lieber, M. M., Bostwick, D. G., Bergstralh, E. J., et al. (1999) Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma. *J. Natl. Cancer Inst.* **91**, 1574–1580.
8. Freedman, M. L., Haiman, C. A., Patterson, N., McDonald, G. J., Tandon, A., Waliszewska, A., et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African American men. *Proc. Natl. Acad. Sci. U S A* **103**, 14068–14073.
9. Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.
10. de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223.
11. Stephens, M., and Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73**, 1162–1169.

12. Thomas, D. C. (2004) *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford.
13. Stram, D. O., Haiman, C. A., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., et al. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55**, 27–36.
14. Pritchard, J. K., and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet*. **69**, 1–14.
15. Stram, D. O. (2005) Software for tag single nucleotide polymorphism selection. *Hum. Genomics* **2**, 144–151.
16. Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet*. **74**, 106–120.
17. Chapman, J. M., Cooper, J. D., Todd, J. A., and Clayton, D. G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered*. **56**, 18–32.

4

Measuring the Effects of Genes and Environment on Complex Traits

Jennifer H. Barrett

Summary

Complex diseases and traits are influenced by a combination of genetic and environmental risk factors, some of which may be known, and many of which are unknown. It is possible to estimate the relative importance of the influence of genes and environment on a trait by studying correlations in the trait in related individuals. Known risk factors can be measured and included in the statistical models to understand disease etiology better. The joint effect of specific genes and environmental exposures can be estimated by measuring these in individuals, not necessarily related, with and without the disease of interest or with a range of trait values. These methods are illustrated by considering two example analyses in detail. The first is an analysis of a study of adolescent twins, quantifying the effect of genes and environment, including measured sun exposure, on the density of nevi. The second is an analysis of a case-control study, examining the joint effect of the *GSTT1* gene and vegetable intake on risk of colorectal cancer.

Key Words: genes; environment; complex traits; variance components; gene-environment interaction; twins; case-control study; statistical models.

Abbreviations: DZ – dizygotic, GST – glutathione S transferase, ICC – intraclass correlation coefficient, MZ – monozygotic

1. Introduction

Most common diseases and clinically relevant traits are influenced by a large number of genetic and environmental risk factors, most of which individually have only a modest effect on the risk of disease or on the trait. Moreover, the

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

manner in which these factors influence disease risk is likely to be complex, so that the joint effect of multiple risk factors may not be predictable from their individual marginal effects. In this chapter some epidemiological and statistical methods for investigating the etiology of such complex or multifactorial diseases or traits, and for furthering understanding of the joint effects of or interaction between risk factors are presented.

Two types of clinically motivated questions in this general field of enquiry are considered, and will be illustrated by one example of each. The first question of interest concerns estimation of the relative contributions of genes and environment on the trait or on disease risk. By measuring the binary disease state or the quantitative trait (the "phenotype") on a sample of related individuals, it is possible to estimate the relative importance of genes and environment, even though these may not have been measured. The second question concerns estimation of the effect on phenotype of specific measured genetic and environmental factors and investigation of their joint mode of action. This can be addressed by measuring phenotype and the specific environmental and genetic factors of interest in individuals, who may or may not be biologically related.

Example 1: Quantifying the overall effects of genes and environment on a quantitative trait—investigating nevus density and sun exposure by studying twins.

A high density of benign melanocytic nevi is an important risk factor for melanoma ***(1)***, and hence there is considerable interest in understanding what determines nevus phenotype. To what extent are nevi determined by genetic factors, and to what extent by environmental risk factors, in particular sun exposure? This question can be addressed by studying twins, using only their nevus phenotype and self-reported sun exposure (***2–4***).

Example 2: Investigating the joint effect of specific genotypes and environmental factors on risk of disease—investigating the joint effect of glutathione-S-transferase (GST) genes and dietary factors on risk of colorectal cancer.

Diet is known to be an important influence on the risk of colorectal cancer. High vegetable consumption has been associated with decreased risk ***(5)***, although not consistently ***(6)***, and high consumption of red meat is associated with increased risk ***(7)***. Many genes, including GSTs, have now been identified that influence the metabolic pathways involved in processing dietary carcinogens or anti-carcinogens. Thus, a question of interest is whether or not an individual's genotype affects the extent to which different dietary factors influence risk. The protective effect of cruciferous vegetables may in part be due to isothiocyanates, an ingredient with a known anti-carcinogenic effect ***(6)***. Here we investigate whether the protective effect afforded by high vegetable consumption depends on *GSTT1* genotype. This question can be addressed by studying unrelated individuals with and without colorectal cancer (cases and controls) from the same population, and measuring genotype and dietary intake in the study participants.

2. Materials

Statistical software is needed for these analyses:

1. Stata Statistical Software Release 9 (College Station, TX: StataCorp, 2005) or a similar general statistical software package is required for both types of analyses presented here.
2. MX software is a matrix algebra interpreter and numerical optimizer suitable for structural equation modelling and other types of statistical analyses. It is written and maintained by Michael Neale and others from the Virginia Commonwealth University, and is freely available to download from the webpage www.vipbg.vcu.edu/∼mx/mxgui/. The web page also includes numerous example scripts, including analyses of twin studies.
3. SOLAR (Sequential Oligogenic Linkage Analysis Routines) is software for genetic variance components analysis. It is written and maintained by John Blangero and colleagues from the Southwest Foundation for Biomedical Research and is freely available to download from www.sfbr.org/solar/.

3. Methods

3.1. Quantifying the Overall Effects of Genes and Environment on a Quantitative Trait—Investigating Nevus Density and Sun Exposure by Studying Twins

Many traits, including nevus density, are correlated within families. By modeling the pattern of correlations between different types of relative-pairs, it is possible to distinguish between correlation due to shared genes and correlation due to shared environment. Hence estimates can be made of the "heritability" of the trait, i.e., the proportion of variability due to genes.

Although other family structures can be used, the twin design is very simple and powerful for such studies. Twin pairs are ascertained from the population, taking no account of trait values or zygosity status, i.e., there is no selection for extreme traits, and both monozygotic (MZ) and dizygotic (DZ) twin pairs are included. The phenotype of interest is measured on each twin in the study (*see* **Note 1**), along with any potential confounding factors, i.e., factors related to both the trait and to potential environmental factors, such as age and sex. If there is also interest in incorporating the effects of known environmental factors (*see* 3.1.3), these exposures are also recorded for each study participant.

It is crucial to the success of this method that the relationship between the twins is accurately known. Zygosity status can be obtained from the twins themselves by self-report, although this is unlikely to have 100% accuracy. If DNA is already being collected from study participants, it is advisable to genotype a small number of highly polymorphic markers across the genome to check zygosity status in same-sex twin pairs (*see* **Note 2**).

3.1.1. Preliminary Analysis

The method employed for the primary analysis of heritability assumes that the phenotype is approximately normally distributed. It is also advisable to remove the effect of any extraneous factors that have a major influence on phenotype (in this case age and sex). Therefore, before addressing the main question it is important

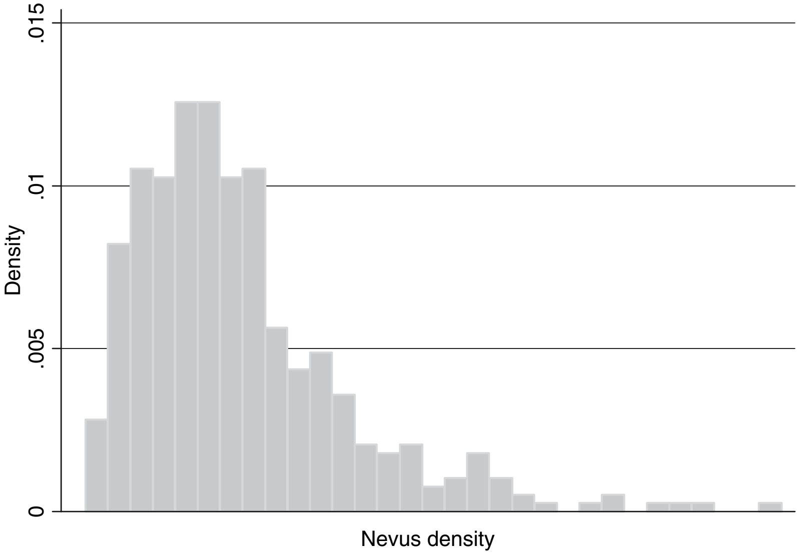


Fig. 1. Histogram showing distribution of nevus densities for 426 adolescent twins from the UK.

to have a proper understanding of the distribution of the trait and its relationship with other variables. The following steps should be carried out:

1. Calculate summary statistics (mean, median, range, standard deviation) of the phenotype for the overall sample.
2. Plot a histogram of the distribution of the phenotype. **Figure 1** shows the distribution of nevus density in 426 adolescent twins from the UK *(3)*. It can be seen that the distribution is highly skewed. As a result of this, the data were log-transformed before further analysis, which resulted in an approximately normal distribution.
3. Regress phenotype on potential confounding variables, e.g., age and sex. This can be done using any standard statistical software package such as Stata (*see* Materials). For a continuous outcome such as nevus density, the regression equation is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

   where y_i is the phenotype for individual *i* (in this case $\log_e$(nevus density)), x_i is a vector of covariates for individual *i*, ε_i is a normally distributed error term with mean 0, and α and β are the coefficients to be estimated (*see* **Note 3**).
4. From the results of the regression model, examine the estimates of the parameters β and their confidence intervals. Any covariates that significantly influence the phenotype should be retained in the model. Other covariates can be included if there is a good a priori reason for doing so. In our example sex was a significant predictor of nevus density (estimated coefficient −0.19, 95% confidence interval (CI) (−0.31, −0.08), $P = 0.001$), indicating that the log nevus density was on average 0.19 units higher in boys than girls. Although age was not a significant predictor of nevus density in this sample, where the age ranged from 10 to 18 years, it was retained in the model, since it is known that nevus density does vary with age *(8)*.
5. Repeat the regression analysis with fewer covariates if necessary and examine the distribution of residuals from the model (the residuals are the differences between the observed phenotypes y_i and the values that are predicted by the model on the basis of the observed covariates x_i and the parameter estimates). Since it is planned to use the residuals from the model in the heritability analysis, it is necessary to check their distribution. Since the raw phenotype data in our example were log-transformed to achieve normality, the distribution of residuals would be likely to be normal, and this was indeed the case. This results in a new "phenotype" suitable for use in the heritability analysis.

3.1.2. Heritability Analysis

The main question posed—*To what extent are nevi determined by genetic and to what extent by environmental risk factors?*—can now be addressed. The idea is to examine the correlation in phenotype between twin pairs. In a twin study there are only two types of relative pairs—DZ twins and MZ twins. The higher the correlations in phenotype are among MZ compared with DZ pairs, the greater the genetic component must be.

1. This can first be explored visually by drawing scatterplots of the phenotypes of one twin from each pair versus their co-twin, separately in DZ pairs and in MZ pairs. **Figure 2** illustrates this for the example dataset. By visual inspection it is clear that the correlation is considerably stronger in MZ pairs.
2. A useful preliminary analysis is to estimate the intraclass correlation coefficient (ICC), measuring the correlations in phenotype between co-twins in MZ and DZ twin pairs, using one-way analysis of variance. The outcome measure is the (transformed, adjusted) phenotype and the classes are defined by the twin-pairs. More formally,

   $$y_{ij} = \alpha + z_i + \varepsilon_{ij}$$

   where y_{ij} is the phenotype for the *j*th twin (j = 1 or 2) in twin pair i, z_i is a random effect for twin pair i representing systematic deviation of that twin-pair from the population mean, and ε_{ij} represents individual-level variation. The random effect z_i is assumed to be normally distributed with mean 0 and variance σ_u^2, and ε_{ij} is assumed to be normally distributed with mean 0 and variance σ_e^2. The ICC is $\sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$, i.e., the proportion of the total variance that is between, rather than within, twin pairs. It can be seen that if there is very little variation within each twin pair (the twins having very similar phenotypes) then σ_e^2 will be small and the ICC close to 1. Conversely, if twins are no more like each other than they are like other individuals, then the twin variable z_i does not explain any of the variance and the ICC is close to 0. In the example study the estimated ICC for DZ twins was 0.61 (95% CI (0.49, 0.72)) and for MZ twins was 0.94 (95% CI (0.91, 0.96)). (*see* **Note 4**).
3. The ICCs for DZ and MZ twin pairs can be compared using a simple test based on Fisher's z-transformation, which transforms a correlation coefficient *r* to a normal statistic:

   $$z(r) = ½\log_e[(1+r)/(1-r)]$$

   The test statistic for the difference in ICCs is

   $$Z = [z(\mathrm{ICCM}) - z(\mathrm{ICCD})]/\sqrt{}(1/(m-2) + 1/(d-2))$$

   where *ICCM* and *ICCD* are the ICCs for MZ and DZ twin pairs and *m* and *d* are the numbers of MZ and DZ twin-pairs respectively ***(9)***. Under the null hypothesis of equal ICCs, Z has an approximately standard normal distribution. In the example study, Z = 7.4, so there is overwhelming evidence that the ICCs are indeed different.

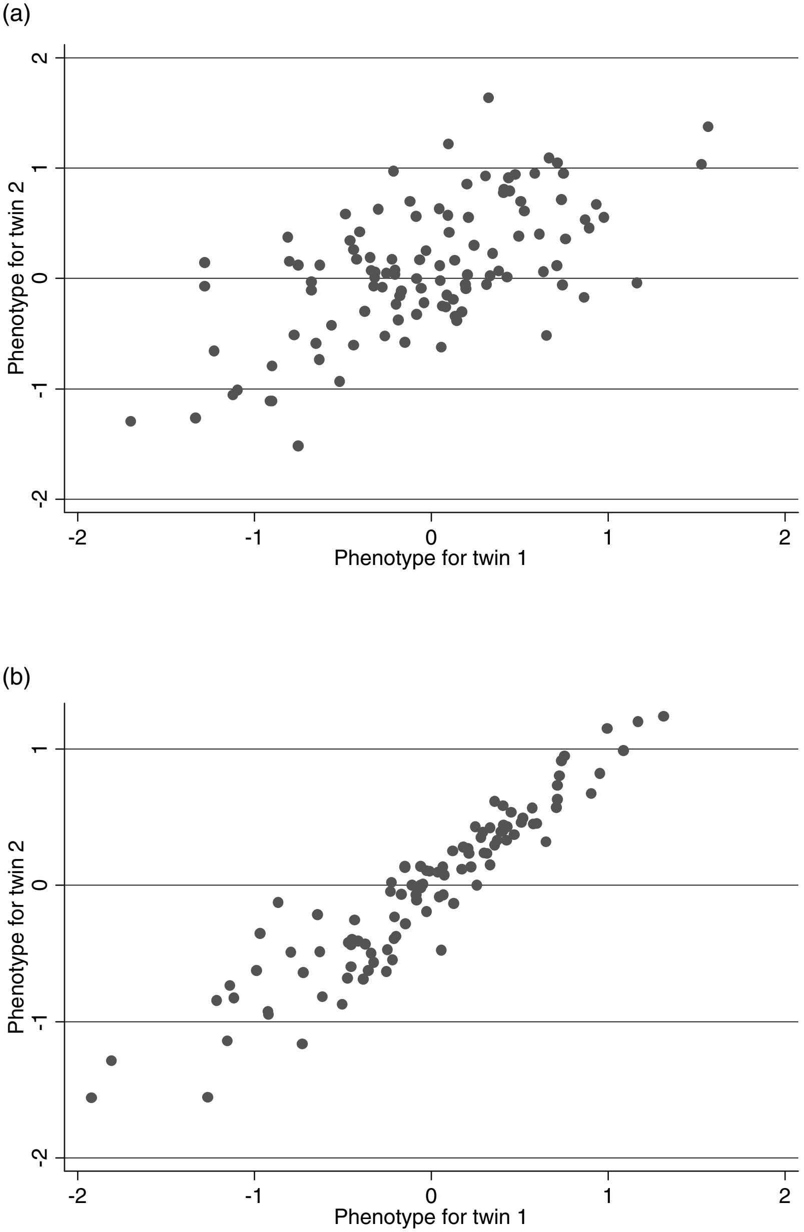


Fig. 2. **a:** Scatterplot demonstrating the correlation in phenotypes of DZ twins based on 110 pairs. **b:** Scatterplot demonstrating higher correlation in phenotypes of MZ twins based on 103 pairs.

4. To estimate heritability, first calculate the variance-covariance matrices for the phenotypes of DZ and MZ twin pairs separately. The matrix for DZ twin pairs from the example data is shown below:

$$D = \begin{pmatrix} 0.3710 & - \\ 0.2285 & 0.3710 \end{pmatrix}$$

The diagonal entries represent the phenotypic variance and the estimates are based on all 220 DZ twins (from 110 pairs) since there is nothing to distinguish the two twins in a pair ***(10)***. The off-diagonal entry is the covariance between the two twins in a pair. The corresponding matrix for the 103 MZ twin pairs in the example data is

$$M = \begin{pmatrix} 0.3605 & - \\ 0.3386 & 0.3605 \end{pmatrix}$$

5. The final step in the analysis is to model the observed variance-covariance matrices by considering the underlying components of variance. Conceptually the phenotype Y, after standardization, can be modelled as a linear combination of three standard normally distributed latent variables, representing additive genetic effects (A), environmental effects shared by co-twins (C), and individual environmental effects which include measurement error (E):

$$\mathrm{Y} = \mathrm{aA} + \mathrm{cC} + \mathrm{eE}$$

The analysis is based on the fact that MZ twins share all their genes and DZ twins share on average half their genes, and this contrast is exploited to estimate the genetic component of variance. It is further assumed that MZ and DZ twin pairs share environmental risk factors to the same extent (*see* **Note 5**). Then the variance-covariance matrices for DZ and MZ twins are predicted ***(11)*** to be

$$D = \begin{pmatrix} a^2 + c^2 + e^2 & - \\ \frac{1}{2}a^2 + c^2 & a^2 + c^2 + e^2 \end{pmatrix}$$

$$M = \begin{pmatrix} a^2 + c^2 + e^2 & - \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{pmatrix}$$

Specialist software such as MX (*see* Materials) can then be used to fit this model to the observed matrices. In the example data set the additive component of variance a^2 is estimated at 65.5% (95% CI (46.9, 89.9)), the environmental component shared between co-twins c^2 is estimated to be 28.5% (95% CI (4.0, 47.3)) and the residual individual level component e^2 6.0% (95% CI (4.4, 8.3)).

3.1.3. Incorporating Measured Environmental Exposure into Heritability Analysis

It is known that higher levels of sun exposure lead to an increase in nevus density ***(12)***. By incorporating measurements of sun exposure into the heritability analysis, the relative importance of the factors contributing to nevus phenotype can be estimated in more detail.

1. In the example study various measures of sun exposure were used, based on responses to questions about time spent on the beach, sunbathing and pursuing other outdoor activities both in the UK and on holidays in hotter countries. A clear relationship with nevus phenotype was found for sun exposure on holiday in countries hotter than the UK ***(2)***. For example, those in the highest quartile of time spent on the beach in hotter countries had a nevus density 15 per m^2 higher than those in the lowest quartile (after adjusting for age, sex, skin type, eye and hair color).
2. How much of the variance due to environmental exposure (c^2 + e^2) is due to (measured) sun exposure? This question can be addressed by including sun exposure in the variance components model. Again specialist software is available for such analyses and we illustrate this using SOLAR (*see* Materials), which allows flexible inclusion of covariates. First, the heritability analysis (*see* 3.1.2) with no covariates is run in SOLAR for comparability, this time using the log transformed nevus densities. The estimates obtained are a^2 = 71.3%, c^2 = 22.6% and e^2 = 6.1%, differing slightly from the estimates in section 3.1.2 because the effects of age and sex have not been removed.
3. Next a model is fitted including sun exposure (time spent on holiday in hot countries) as a covariate. By comparing the results of these analyses, estimates can be derived of the proportion of the variance due to environmental exposure that is explained by measured sun exposure.

   SOLAR reports the proportion of total variability in the trait that is explained by the covariate(s), estimates of *residual* heritability, which is the heritability of the trait after removing the variability explained by the covariate, and similar measures for the shared and individual environmental components. In this model, it is found that 8.4% of the variance is explained by this measure of sun exposure. The *residual* heritability rises to over 79.5%, indicating that an even greater proportion of the *remaining* variance is genetic. The residual estimates of common and individual environment are 13.7% and 6.8% respectively. These can be expressed as the proportion of the total variance by scaling as follows:

$$a^2 = 79.5\% \times (100 - 8.4)/100 = 72.9\%$$

$$c^2 = 13.7\% \times (100 - 8.4)/100 = 12.5\%$$

$$e^2 = 6.8\% \times (100 - 8.4)/100 = 6.2\%$$

the remaining 8.4% being due to measured sun exposure. Inclusion of the sun exposure measure has reduced the unaccounted for shared environmental exposure from 22.6% to 12.5%, and had little impact on the estimates of genetic and individual environmental components of variance. This is because sun exposure is an environmental rather than genetic factor, and the time spent on holiday abroad is likely to represent exposure shared by co-twins, all of whom were aged 10–18 years and probably holiday together. In summary, between one third and one half of the variance due to common environmental exposure has been explained by this simple measure of sun exposure.

3.2. Investigating the Joint Effect of Specific Genotypes and Environmental Factors on Risk of Disease—Investigating the Joint Effect of GST Genes and Dietary Factors on Risk of Colorectal Cancer

There are many examples of known environmental risk (or protective) factors for disease where it would be of value to know whether or not the associated risk was influenced by the individual's genotype. This could ultimately be important for public health measures but it can also provide information about the etiological process. For a specific environmental exposure E and genotype G, such questions can be addressed by comparing the risk associated with E in subjects with genotype G to the risk in subjects without the G genotype.

Other designs are possible, such as using unaffected family members as controls, but a simple design is to conduct a case-control study. Cases with the disease (preferably incident) are ascertained and compared with unrelated controls from the same population, possibly matched for important potential confounding factors such as age and sex. Both E and G are measured on each study participant, along with any potential confounding factors i.e. factors related to both the disease and to E.

3.2.1. Preliminary analysis

1. Environmental exposure may be measured on an interval scale or as a binary or other categorical variable. In this example usual consumption of vegetables was measured using a food frequency questionnaire and also a simple cross-check question, which asked about average non-itemized vegetable consumption ***(13)***. Such data are subject to considerable measurement error, but are nonetheless able to distinguish between low and high consumers. A variable is often created based on the quantiles of the distribution in controls, and in this example approximate tertiles of the cross-check question were used, creating the categories of low, medium and high vegetable consumption.
2. As a quality control measure, the genotype distribution in controls should be tested for consistency with Hardy-Weinberg equilibrium. With data only available on

unrelated individuals this is one of the only statistical checks for genotype error that can be performed. Although other factors can also give rise to departure from the expected distribution, genotype error should be investigated as one of the most likely explanations ***(14)***.

3. The main effect of G on disease can be examined by drawing up a contingency table of disease status by genotype. Odds ratios can be calculated comparing the risk of disease in each of the other genotype categories compared with homozygotes for the most common allele. Except in the case of single nucleotide polymorphisms, there may be too many genotypes to consider in this way, but if prior knowledge permits then it may be possible to combine genotypes into a small number of groups. In this example, the *GSTT1* polymorphism investigated is related to function, so that homozygotes for the polymorphism are predicted to have deficient phenotype (complete loss of enzyme function), those with no copy of the polymorphism have fast/active phenotype and those with one copy have intermediate phenotype ***(15)***. In the example study the cases and controls were matched for age and sex, so the odds ratios were calculated from conditional logistic regression analysis of the matched pairs. This can be done using any standard statistical software package such as Stata (*see* Materials). No effect of *GSTT1* genotype on colorectal cancer risk was observed (odds ratio (OR) 0.7 (95% CI (0.5, 1.0)) for heterozygotes and OR 1.1 (95% CI (0.7, 1.6)) for those with predicted deficient phenotype compared with those with predicted fast phenotype).
4. The effect of E on disease risk can similarly be examined. In this case vegetable consumption showed weak evidence of an effect on risk of colorectal cancer, with reduced risk for those in the top two tertiles (OR 0.6, 95% CI (0.4, 0.9) for the intermediate group and OR 0.7 (0.5, 1.0) for the high consumers) compared with the low consumers ***(13)***.
5. Association between disease status and also E and G and any potential confounding variables (such as age and sex) can be examined to decide what confounders should be considered in the analysis. In this example, the study design included matching for age and sex, but other potential confounders included smoking history, body mass index, and other dietary risk factors.

3.2.2. Joint Effect on Risk of Genetic and Environmental Risk Factors

On the basis of the analysis of main effects, it may be decided to group categorical variables into smaller numbers of categories to evaluate joint effects. In this case vegetable intake was categorized into just two groups, low versus intermediate or high, since the last two groups show a similar reduction in risk. The *GSTT1* genotype is still considered as three categories.

1. One way to look at combined effects is to compare individuals with each combination of risk factors with the same baseline category. Here we take the baseline category of *GSTT1*-deficient genotype and low vegetable consumption,

and estimate the relative risk associated with each other combination of risk factors by including these as a 6-level factor in a logistic regression model (**Table 1**). Because the data consist of case-control pairs, matched for age and sex, conditional logistic regression was used. It can be seen that compared with this baseline all other categories are at lower risk of colorectal cancer, with similar estimated odds ratios of around 0.3. Those at highest risk of colorectal cancer are thus individuals with the combination of deficient genotype and low vegetable intake.

2. Does the effect on risk of E differ depending on the presence/absence or value of G? To examine this, the data can be stratified by genotype and within each stratum disease status regressed on the environmental exposure. In this example, the OR for high/intermediate vegetable consumption compared with low is 0.26 (95% CI (0.12, 0.58)) in the deficient genotype group, 0.62 (0.40, 0.96) in the intermediate group and 1.39 (0.81, 2.37) in the fast group. Thus, the protective effect of vegetable consumption is seen most strongly in those with deficient genotype and not at all in those with the fast genotype.
3. To test whether these differences are statistically significant, a model can be estimated including G, E, and the interaction between them (*see* **Note 6**). This additional term or terms allow the effect of E to differ across levels of G. Significance of the interaction is assessed by carrying out a likelihood ratio test comparing the model with interaction with the model excluding interaction. In the example data there is clear evidence for interaction ($P = 0.006$) (*see* **Note 7**).

In summary, the analysis shows that the protective effect of vegetable intake on risk depends on genotype. In those with fast genotype, low vegetable consumption is not associated with increased risk, whereas those with low intake and deficient genotype are at the greatest risk.

Table 1
Combined effect of G (*GSTT1* genotype) and E (vegetable consumption) on risk of colorectal cancer

GSTT1 Genotype	Vegetable Consumption	OR (95% CI)
Deficient	Low	1 (baseline)
Intermediate	Low	0.37 (0.15, 0.87)
Fast	Low	0.26 (0.10, 0.66)
Deficient	Intermediate or high	0.25 (0.10, 0.62)
Intermediate	Intermediate or high	0.23 (0.10, 0.53)
Fast	Intermediate or high	0.34 (0.15, 0.78)

4. Notes

1. There are many important ethical and methodological considerations in conducting epidemiological studies that are beyond the scope of this chapter. Here we assume for example that variables are measured as accurately as possible and that care is taken to minimize potential sources of bias.
2. Twin pairs of opposite sex are clearly DZ. For all same-sex twin pairs with alleles identical (by state) at each genotyped marker, the probability of dizygosity can be calculated based on the observed genotype using estimated population allele frequencies. If sufficiently polymorphic, only a small number of markers (*6–10*) are needed to classify twins concordant at these markers as MZ with high probability ($P > 0.999$). Any twin pair discordant at any genotyped marker would be classed as DZ by definition, but it is important to bear in mind the possibility of genotype error. It is thus better to use a small set of highly reliable markers than to include any less reliable markers in this calculation. If twins seem to be discordant at one marker only this should be checked carefully.
3. This is a simplified model in that, if all twins are used, it ignores the correlations between twins in a family. This will lead to an underestimate of the standard errors of the estimated coefficients. However, for the purposes of this analysis (adjusting for covariates) the model is adequate. Alternatives, which would produce similar adjusted values, would be to assess statistical significance using a random effects model allowing for clustering within families, or to analyse one twin selected at random from each pair.
4. It might have been expected that correlation would be measured using the standard Pearson correlation coefficient. However, twins within a pair are not distinguishable in any natural way. If we labeled the twins *twin1* and *twin2*, the labelling would be entirely arbitrary (with respect to phenotype). Hence ICCs, which do not impose this ordering, provide a preferable measure.
5. The assumption that DZ twins share environmental exposures to the same extent as MZ pairs is of course open to question, but it is reasonable to assume that differences will generally be small, certainly compared with the difference in the proportion of genes shared.
6. Statistical interaction is not equivalent to biological interaction. One important point is that statistical interaction is model and scale dependent. With simple binary risk factors G and E, statistical interaction means that the data are not consistent with a particular model that predicts the joint effects on risk of G and E. Most commonly, models are used based on multiplicative effects, as has been done here by using a logistic regression model. Statistical interaction simply indicates that the data are not consistent with a multiplicative joint effect. Lack of statistical interaction is still consistent with a higher risk to those with both G and E and certainly does not rule out biological interaction. For example, if G alone doubles disease risk and E alone trebles risk, then under a multiplicative model with no interaction the combined effect would be to increase risk 6-fold.

Statistical interaction simply implies the risk is greater than (or alternatively less than) this. Particular care must be taken in interpreting evidence of interaction when E is measured on a continuous scale, since whether or not there is statistical interaction may depend on the scale of measurement, e.g., whether or not the data are log-transformed.

7. A *P*-value has been quoted here, but in practice the interpretation of results depends very much on the context, since studies of gene-environment interaction are often conducted in the context of multiple testing. Bonferroni correction is usually inappropriate, both because the tests may not be independent and because the number of tests is not well specified (even if 10 hypotheses are tested in one study, another 10 equally likely hypotheses may be tested later). The interpretation of results depends on the prior probability of the hypothesis being true and on the study power. These considerations have been put into a more formal framework, from which estimates can be derived of the probability that the result is a true finding ***(16)***.

References

1. Swerdlow, A. J., English, J., MacKie, R. M., O'Doherty, C. J., Hunter, J. A., Clark, J., et al. (1986) Benign melanocytic naevi as a risk factor for malignant melanoma. *Br. Med. J. (Clin Res Ed.)* **292**, 1555–1559.
2. Wachsmuth, R. C., Turner, F., Barrett, J. H., Gaut, R., Randerson-Moor, J. A., Bishop, D. T., et al. (2005) The effect of sun exposure in determining nevus density in UK adolescent twins. *J. Invest. Dermatol.* **124**, 56–62.
3. Wachsmuth, R., Gaut, R., Barrett, J. H., Saunders, C. L., Randerson-Moor, J. A., Eldridge, A., et al. (2001) Heritability and gene–environment interactions for melanocytic nevus density examined in a UK adolescent twin study. *J. Invest. Dermatol.* **117**, 348–352.
4. Zhu, G., Duffy, D. L., Eldridge, A., Grace, M., Mayne, C., O'Gorman, L., et al. (1999) A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: a maximum-likelihood combined linkage and association analysis in twins and their sibs. *Am. J. Hum. Genet.* **65**, 483–492.
5. Potter, J. D. (1999) Colorectal cancer: molecules and populations. *J. Natl Cancer Inst.* **91**, 916–932.
6. Michels, K. B., Edward, G., Joshipura, K. J., Rosner, B. A., Stampfer, M. J., Fuchs, C. S., et al. (2000) Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. *J. Natl. Cancer Inst.* **92**, 1740–1752.
7. Norat, T., Bingham, S., Ferrari, P., Slimani, N., Jenab, M., Mazuir, M., et al. (2005) Meat, fish, and colorectal cancer risk: the european prospective investigation into cancer and nutrition. *J. Natl. Cancer Inst.* **97**, 906–916.
8. Gallagher, R. P., and McLean, D. I. (1995) The epidemiology of acquired melanocytic nevi. A brief review. *Dermatol. Clin.* **13**, 595–603.

9. Donner, A. (1986) A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *Int. Statist. Rev.* **54**, 67–82.
10. Sham, P. (1998) *Statistics in Human Genetics*. Arnold, London, p.242.
11. Eaves, L. J. (1977) Inferring the causes of human variation. *J. Royal Statist. Soc. Series A*, **140**, 324–355.
12. English, D. R., Milne, E., and Simpson, J. A. (2006) Ultraviolet radiation at places of residence and the development of melanocytic nevi in children (Australia). *Cancer Causes Control* **17**, 103–107.
13. Turner, F., Smith, G., Sachse, C., Lightfoot, T., Garner, R. C., Wolf, C. R., et al. (2004) Vegetable, fruit and meat consumption and potential risk modifying genes in relation to colorectal cancer. *Int. J. Cancer* **112**, 259–264.
14. Wittke-Thompson, J. K., Pluzhnikov, A., and Cox. N. J. (2005) Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967–986.
15. Sachse, C., Smith, G., Wilkie, M., Barrett, J., Waxman, R., Sullivan, F., et al (2002) A pharmacogenetic study to investigate the role of dietary carcinogens in the etiology of colorectal cancer. *Carcinogenesis* **23**, 1839–1849.
16. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, 434–442.

5

Microarrays—Planning Your Experiment

Jean Yee Hwa Yang

Summary

The rapid increase in the use of microarray studies has generated many questions on how to plan and design experiments that will effectively utilize this technology. Investigators often require answers to questions relating to microarray platforms, RNA samples, options for replication, allocation of samples to arrays, sample sizes, appropriate downstream analysis, and many others. Careful consideration of these issues is critical to ensure the efficiency and reliability of the actual microarray experiments, and will assist in enhancing interpretability of the experimental results.

Key Words: Experimental design; microarray; gene expression; probe design; replication; randomization.

1. Introduction

Good experimental design in microarray studies simplifies analysis and enhances interpretation of data. Various considerations go into the planning of an effective experiment. In the last few years, many of the publications on experimental strategies have focused on the identification of an efficient design. While this is still an important component, there are broader considerations. In this chapter, the various aspects involved in planning a successful microarray experiment will be described.

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

2. Materials

When planning an experiment, a number of general issues need to be identified to help translate the overall biological questions into a more defined and appropriate statistical framework. Some of these include

1. *Aim of the experiments*: This refers to the biological question of interest. Given that microarrays are utilized in a wide variety of contexts, this consideration can be very specific or general. For example, a researcher can study a focused hypothesis such as identifying differences between wild-type and mutant mice. Alternatively, a research can perform an experiment with a very general aim in mind, such as profiling gene expression from a collection of clinical patients and then attempt to generate a possible hypothesis from the data. Regardless of the type of aim, it is important to consider how these aims will contribute to understanding further the long-term goal of the research (*see* **Note 1** and **Chapters 6–9** for examples of different microarray studies including experimental designs).
2. *Main comparison*: Researchers often wish to investigate multiple questions in a single experiment. Given that each question has implications for experimental design and downstream analysis, researchers need to clarify the specific questions being asked and subsequently, the most important question or comparisons. One such example is the intention to derive simultaneously a classification rule that best predicts survival outcome of cancer patients as well as to identify a collection of biomarkers that best distinguish the survival outcome.
3. *Resources*: This determines the size and scale of the experiment and is predominantly dependent on the number of samples available, the amount of mRNA and ultimately funding for the experiment.
4. *Previous experiments*: It is not always clear whether changes between mRNA are detectable by microarray experiments with many of the genome-wide profiling studies. For such situations pilot studies to determine the feasibility of a larger study should be considered.
5. *Verification method*: It is possible to overlook the limitation of the arrays and the way in which these data contribute to the overall research goal as researchers get caught up in the novelty of a technology. As microarray is still a relatively new experimental approach, large variability in results can be expected. Consideration needs to be given to method**s** required to verify results obtained from microarrays.

3. Methods

There are two main aspects to planning a successful microarray experiment. The first concerns the actual design or selection of the microarray. This refers to the choice of DNA probes to print onto the solid substrate, e.g., a membrane, glass slide, or silicon chip, and where they are to be printed. The second aspect concerns the planning and design of the actual hybridizations. This often refers to the allocation of target samples to the microarrays, the nature, and the number of replications required **(***see* **Note 2**).

3.1. Selection of Microarray

Selection of an appropriate microarray platform is an essential component in planning microarray experiments. There are many recent publications on experimental designs and downstream analysis of gene expression data. However, the actual design and the effects of the microarray platform are often ignored or not considered. As scientists propose more complex array based experiments, it is essential that serious consideration is given to this aspect of experimental design during planning.

The first question that any investigator faces when selecting a microarray platform is whether to invest in designing in-house arrays or to purchase prefabricated arrays. The purchase of pre-fabricated arrays refers to purchasing arrays produced by commercial companies and non-commercial facilities. Naturally, this will depend on the availability of an appropriate and satisfactory platform and the size of the experiment. Pre-fabricated arrays are generally preferred for genome-wide scans of popular genomes such as human, mouse or rat. Section 3.1.2 will discuss the selection of an appropriate platform and libraries from the many choices available. The demand for in-house design and production of boutique and custom arrays often comes from investigators who wish to focus on specific biological processes and are not as interested in a global genetic perspective. For a study of vasculogenesis, for example, a few hundred genes related to vasculogenesis can be selected with a boutique array produced with just those clones. Alternatively, there remain many genomes that are not fully sequenced or have no pre-manufactured arrays available. In these cases, the investigator will need to plan and design in-house arrays.

3.1.1. In-house Probe Design

The two most widely used custom build microarrays are spotted cDNA and spotted long oligonucleotide (oligo) arrays. A summary of key issues to consider during the design of oligo probes will be discussed in this section.

1. Selection of *gene-representative sequences* from a given *gene collection.*
   The determination of the probe sequences to be printed is an important and specialized bioinformatics task that attempts to identify optimal exon probes (the best 60–70 base pairs) that characterize the genes of interest. Properties that make a good oligo-probe are

   a. *Sensitive*: referring to a strong signal for the complementary target. These probes will have no secondary structure in the probe or target and are located near the 3’ end as preferred by oligo dT.

   b. *Specific*: implies a probe that returns weak signals for non-targets. This means there is no cross-hybridization to other targets.

   c. *Isothermal*: the probe behaves similarly under the hybridization conditions of the microarray experiment such as temperature, salt, and formamide concentration.

   Examples of algorithms that perform the selections of sequences include OligoPicker ***(1)***, ArrayOligoSelector ***(2)***, and OligoArray ***(3)***. The reader is referred to these articles for more details on probe selection.

2. *Location and number* of the probes.
   Randomization is often needed to avoid systematic spatial biases of the probes.
3. *Control* probes.
   A well-designed set of controls allows detailed assessment of the performance of any given hybridization, including sensitivity, specificity, dynamic range, normalization, linearity, and allows various biases to be explored. A recently developed open-source oligonucleotide set "Exonic Evidence-Based Oligonucleotide Chip (MEEBO and HEEBO)" as a collaborative effort between researchers at University of California at San Francisco (UCSF), Stanford, Rockefeller, Basel, and the Stowers Institute has produced an unprecedented large number of controls probes. Some examples from the Mouse EEBO are

   a. *positive* controls that show a strong signal able to be used as "landing lights" to assist in the image analysis.
   b. *negative* controls, such as blank spots or spots with cDNA from very different species that provide an estimate of background signal.
   c. *doped* controls are probes that recognize non-mouse sequences that can be spiked into RNA samples. These can be used as normalization controls with appropriate spiked-in mixture.
   d. *tiling* controls to assess effects between observed hybridization intensities and distance from the 3' end.
   e. *mismatch* controls—used to help fine tune hybridization conditions.

3.1.2. Which platform to purchase?

In this section, we will discuss issues to consider when selecting prefabricated array platforms. As mentioned before, arrays are available from commercial companies as well as various core facilities set up by research and academic institutes. We will refer to all of these as "array providers" regardless of whether they are providing commercial or non-commercial platforms. Broadly speaking there are four main types of microarray platforms. These are

1. Short oligonucleotide (with 25 base pairs) arrays, e.g., Affymetrix.
2. Two-color cDNA spotted arrays.
3. Two-color long-oligonucleotide spotted arrays. The production of this platform is very similar to the two-color cDNA arrays the main differences is in the length of the sequence (60–75 bp) spotted on the arrays.
4. Beaded arrays, e.g., Illumina.

When selecting a pre-fabricated array, the probe design decision is generally pre-determined by the array provider. Considerations include

1. *Probe of interest*. A key consideration is whether the genes or gene sets of interest are included in the arrays. For two-color cDNA and long-oligonucleotide arrays, there are collections of different libraries and probe sets that can be selected. Examples from the cDNA arrays include libraries generated by the Riken consortium, the NIA group. Examples from the long-oligo librarys include array-ready oligo sets designed and produced by commercial companies such as Operon, Agilent, and Illumina.

   For users who are interested in comparing the similarities and differences between these libraries, a useful web tool is Resourcerer ***(4)*** (http://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl), which provides annotations based on the TIGR Gene Indices for commonly available microarray resources, including widely used clone sets. This tool also allows comparisons between resources from the same species. A recent study also provides a detailed comparison of gene coverage of mouse oligonucleotide microarray platforms ***(5)***.
2. *Controls*. Companies that produce short-oligonucleotide and beaded arrays have built-in a series of quality and normalization controls lists of which are available to researchers. For two color array technologies, the number and types of quality and normalization controls varies greatly between array providers. It is important to obtain similar information on control probes from the array providers. Furthermore, one needs to bear in mind the downstream analysis when evaluating whether these controls are adequate for the experiment. For example, many commonly used normalization methods such as print-tip loess rely on the assumption that the majority of genes are not differentially expressed between mRNA samples. However, if the researcher expects a large number of changes in the experiment, alternative normalization methods based on external controls will be required. In this situation, it is important to ensure that normalization controls such as spike in controls or microarray sample pool ***(6)*** are included in the array design (*see* **Notes 3**, **4**).
3. *Print-run quality*. Just like any chemical reagent or paint, there are potential differences between batches of arrays from different print-runs. Researchers could enquire from individual array providers any print-run quality controls that are performed routinely at the facilities and any supplementary print-run information that could aid in the downstream analysis. For example, the UCSF Shared Microarray Facilities routinely perform 9mer hybridizations on selected arrays in every print-run. This allows them to estimate the number of probes (often very small) that failed to print and are able to provide downstream users a list of probe IDs that are problematic on every print-run. This information will facilitate the identification that a low observed intensity of a particular probe is due to problematic print rather than low expressed transcript levels.
4. *Cost of hybridization*. While the cost of the arrays typically refers to the actual cost of purchasing the arrays, the real cost is determined by the cost involved in producing successful array hybridizations. The *reproducibility of hybridization* and

the cost of all the failed hybridizations should be considered in project costing. For many clinical studies, requirements for recruiting large, clinically well-characterized subject cohorts in addition to difficulties inherent in obtaining suitable tissues for the study translate into a higher sample cost compared to the cost of an array. In these situations, the platform of choice should be primarily the most reproducible platforms and the cost of array is less relevant.

5. *Extensibility*. Changes in platforms between different stages of experiments are best avoided, highlighting the importance of the continued availability of the platform of choice. In addition, thought should be given to potential batch effect, e.g., between different print-runs, the changes in the probes between print-runs. This information should be obtained from the array provider prior to the experiment.
6. *Data integration*. The availability of many public data repositories such as ArrayExpress ***(7)*** and NCBI's Gene Expression Omnibus ***(8,9)*** makes it possible to identify other studies that might complement the proposed research. Many statistical methods are currently under development to integrate experiments from different protocols obtained by multiple groups ***(10,11)***. Individual researchers may want to increase their research capacity by considering integrating external studies with their own. To assist in data integration by reducing the potential variation due to combining data from different platforms, the investigator may decide to use the same platform as that in the external studies.

3.2. Planning Hybridizations

Raw data extracted from various array technologies can be broadly split into two groups; single color and two-color arrays. Single color arrays refer to array technologies where only one sample is hybridized on each array, e.g., Affymetrix Gene Chip. Two-color arrays refer to technologies where two or more mRNA samples are hybridized on each. Examples of these include two-color cDNA and two-color long-oligo arrays. Both these technologies share many similar experimental design issues. Readers are referred to the extensive classical literature on experimental design (***12–15***). These books discuss the general principal behind randomization, replication and local control based on agricultural and scientific experimentation. In **sections 3.2.1–3.2.3** a summary of these issues in microarray experiments will be provided. Most of these discussions apply equally to both single color and two-color technologies.

3.2.1. Randomization

The primary objective of randomization is the avoidance of bias or systematic error. The following experimental design could be carried out to identify differential expression between long and short term cancer survivors:

a. tissues for long-term survivors arrive in the laboratory in January and the tissues for short term survivor are sent from the hospital in July; and
b. arrays are run on the tissues samples as they arrive.

Notice there is a complete confounding between survival status and "hybridization time." It is often observed that microarray experimental conditions can leave a strong "global signature" in the resulting expression data. That is, we can easily observe or distinguish, e.g., via cluster analysis, arrays processed or hybridized on different days or batches. Therefore, in the example just described, it will be difficult to distinguish whether observed changes are due to different processing time or the survival status. Proper randomization is required to avoid this type of situation.

3.2.2. Replications

As an individual microarray generates expression values for tens and thousands of genes, it is easy to forget that there is no replication associated with any individual array. That is, there is only one measurement per gene. So why is it essential to replicate slides? The simple answer is that replication reduces variability in summary statistics and permits the use of formal statistical methods.

In essence, replication permits averaging, and averages of independent and identically distributed quantities have less variability than their individual components. For example, at an individual gene level, a gene which seems to be 4-fold differentially expressed in one hybridization experiment may appear to have a 2.3-fold change in a second independent hybridization and 3-fold in a third. To verify that differentially expressed genes between two samples of mRNA are real observations requires replication ***(16)***. This will allow the estimate of the variance of the log-ratios across slides to be calculated. Figure 1 shows three scatter plots of $M = \log_2 (beta7+/beta7-)$ and $A= (\log_2 \sqrt{(beta7+)\cdot(beta7-)}$ values averaged across replicate slides. The data are from a study ***(17)*** aimed to identify differentially expressed genes between the beta7+ and beta7– memory T-helper cells. Each of the replicate slides involves mRNA from different patients, co-hybridized with beta7+ and beta 7– mRNA samples. The figure shows that as the sample size increases (from one to six), the cloud of points about the horizontal axis shrinks. This makes it easier to distinguish real changes from the random variation about zero and demonstrates that replication is a highly desirable feature in planning a microarray experiment. The next question is to determine the number and types of replication.

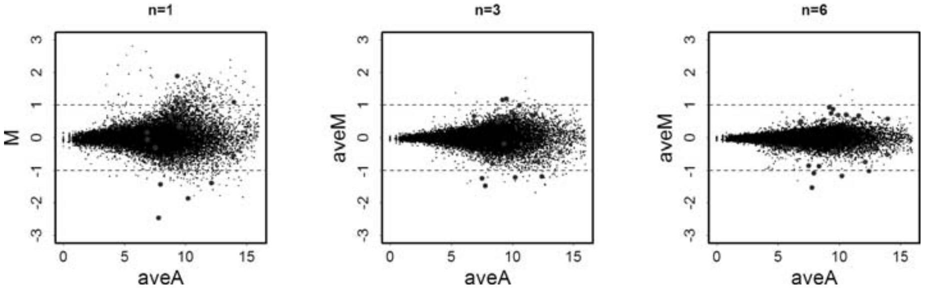


Fig. 1. MA-plot. Scatter plots of average log-ratios $M = \log_2 (beta7+/beta7-)$ averaged across replicate slides, against overall intensities $A = \log_2 \sqrt{(beta7+) \cdot (beta7-)}$. The blue spots correspond to probes that are found to be differentially expressed between the two mRNA sources. The number of replicate slides shown are n = 1, 3, and 6. *See* **Chapter 6** for further discussion on MA-plots.

3.2.2.1. Types of Replication

There are many different types or levels of replication. These will ultimately determine the degree of generalization or conclusions that can be made from the experiment. We can broadly classify the types of replicate into two classes: technical and biological replicates.

1. *Technical replicates.* The term *technical replicate* is used to denote replicate slides made with target RNA from the same preparation. For example, mRNA from the same mouse is taken and then processed, labeled, and finally hybridized on two different arrays (chips). The results from these two arrays are known as technical replicates as the data are based on the same mouse. There are different degrees of technical replications. For example, an extreme form of replication will be having the mRNA undergo the same level of processing at all different stages and only separate the mixture just prior to hybridization. A less extreme form of technical replication could be obtaining mRNA from the same mouse, letting it undergo different amplifications, and processing steps before hybridizing on different chips.
2. *Biological replicates.* The term *biological replicate* refers to hybridizations involving RNA from independent preparations of different samples from the same tissue or cell line. The term *biological replicate* may also refer to replicate slides using target RNA from preparations from different organisms or different versions of a cell line. This illustrates that there are many different levels of biological replicates and these impact of the generality of the experimental results obtained. For example, if a conclusion applicable to all mice of a certain inbred strain is sought, experiments involving multiple mice, preferably random samples of such mice must be performed. Extrapolating to all mice of that strain from results on

a single mouse, even using multiple mRNA extractions, has well-known dangers associated with it ***(18)***.

In general, an experimenter will want to use biological replicates to validate the generality of conclusions and technical replicates to reduce the variability in these conclusions. However, this also needs to be weighted against the cost of the arrays. For example, if you are considering more costly platform such as Affymetrix and using these to perform experiments involves humans, biological replicates are more relevant to the experimental aims than the technical replicates.

3.2.2.2. Number of Replicates (Sample Size)

Given the importance of replication, and having chosen a form of replication suited to the experiment under consideration, an important practical issue is to determine the sample size, or the number of slides to use. A tradition power calculation requires the experimenter to state the variance of individual measurement - the magnitude of the effect to be detected, the acceptable false positive rate and the desired power or the probability of detecting an effect of the specified magnitude (or greater). For microarray studies, the calculation of power is considerably more complex as there are tens of thousands or probe sets and the signal and variance of these probe sets are varied. This is still an active field of statistical research and there have recently been a few proposed methods for samples size determination using statistics to deal with the problem of multiple testing or drawing on pre-existing data for variability estimation (***19–21***). The reader is also referred to ***(22)*** for a more detailed discussion on what can be done for sample sizes determination in the microarray context.

3.2.2.3. Pooled mRNA Versus Unpooled mRNA

An issue closely associated with replication is the discussion on pros and cons relating to pooling mRNA samples before hybridization. We will discuss this issues in two contexts, the first where pooling is not necessary and the second where pooling is necessary.

If one assumes that pooling of mRNA is not necessary, is it nevertheless desirable? Let us suppose that four treated mice and four control mice are to be used in an experiment, and that each mouse would provide sufficient mRNA for a single hybridization. Should the experimenter pool RNA from the four treated mice, do likewise with the control mice, and then carry out the experiment four times using sub-samples of the pooled mRNA? Alternatively, should the experimenter make four separate treatment-control comparisons, and then average the resulting log-ratios? The main argument for pooling mRNA samples is the ability to obtain more precise results with fewer chips and therefore reduce the cost of the experiment. This assumes that the cost of the chips is much greater than the cost per sample, which is often the case in animal studies but may not be so in human studies. Many human or clinical studies involve very high costs in patient recruitment; therefore researchers may need to trade off the cost of arrays versus the cost of mRNA samples. Recently, an experimental study ***(23,24)*** was designed to evaluate the utility of pooling and the impact on identifying differentially expressed genes. The study recommended pooling for small experiments with fewer than three arrays in each condition. However, this study did not find any significant improvement for large experiments and concluded that the "potential benefits from pooling do not outweigh the price paid for loss of individual specific information" ***(24)***.

In other situations it is necessary to pool mRNA from a number of similar sources, e.g., mouse embryos, in order to have sufficient amounts to carry out a single hybridization. In such cases, one needs to assess the possible drawback against possible alternatives, e.g., amplification.

3.2.3. Local Controls (Blocking)

Local controls or blocking refers to arranging experimental units into clusters or blocks in an attempt to improve the comparison of treatments. In the microarray context, we refer to arrangement of the mRNA samples in relationship to various hybridization conditions such as: time of labeling and time of hybridization amongst others. For example, we have discussed in section 3.2.1 that it is not appropriate to hybridize tissues for long-term survivors during the month of January and then samples for the short-term survivors in July. If we know the total number of samples, say, 50, with 25 in each survival group and that the facilities can process 10 samples a day, a better design becomes possible. For example, five samples from the long-term survival group and five samples from the short-term survival group could be processed on the same day. The principal behind this type of experimental design is to group the tissues materials in such a way that unique features associated with the day-to-day processing of arrays are shared equally among the long and short term survivors. This is another way to avoid confounding between hybridization day and survivor status. One can see that this can be extended to different stages of mRNA processing such as tissue extraction, amplification and labeling. For more complex experimental design options, we refer readers to the classical references on experimental design (***12–15***).

3.2.4. Variability Associated with Design Choices for Two-Color Arrays

Most of the discussion so far has applied to all the most widely used types of microarrays. However, the choice of direct versus indirect hybridization is unique to two-color arrays, and so there is some difference to classical experimental design. This is because the two-color arrays are inherently comparative in nature and that actual gene expressions of interest are never measured directly. Therefore, a key component of design issue with two-color spotted microarrays is the decision between using *direct* rather than *indirect* comparisons; that is, comparisons *within* slides rather *between* the slides.

In many cases, given the nature of the experiment and the material available, one design would stand out as an obvious choice. For example, if cells treated with different drugs are to be compared with untreated cells then the appropriate design is clearly one where the untreated cells become a de facto reference, and all hybridizations involve comparison between treated and untreated cells. In another example, suppose that we have collected a large number of tumor samples from patients. If the scientific focus of the experiment is on discovering tumor subtypes ***(25)***, then the design involving comparisons between all the different tumor samples and a common reference RNA is a natural choice. In both cases, the choice follows from the aim of the study, with statistical efficiency considerations playing only a small role. However, with many other experiments, there are a number of suitable choices and some criteria are needed to select one from the set of possibilities. Many of these design choices are discussed in (***22***,***26–29***) where ideas from optimal experimental design are used to select the most efficient approach from a collection of possibilities.

The main ideas presented in these papers are comparing variability associated with different types of experimental design. The simplest case involves treatment T versus control C. The terms *treatment* and *control* are used broadly to include comparisons between the cells or tissues under study and the normal or untreated cells or tissues, e.g., drug versus untreated, wild-type versus mutant (including knock-out or transgenic), or two different tissues (tumor versus normal). Let us compare two simple designs involving two microarrays comparing mRNA T and C. The first involves a direct comparison and the second involves an indirect comparison with a common reference. In a direct comparison, the two samples are co-hybridized together on two slides. For a typical gene, we obtain two independent estimates of the log-ratios (M = log(T/C)). If the variance associated with one measurement is σ^2, then the variance of the average of two independent measurements is $\sigma^2/2$. On the other hand, in an indirect comparison, the two samples will be hybridized on two different slides with a common reference Ref and the log-ratios for a typical genes would be log(T/C) = log(T/Ref) – log(C/Ref). In this case, the variance of the difference of two independent log-ratios is $2\sigma^2$. Therefore, when comparing these two experimental choices holding the number of arrays equal, one would prefer the direct comparison, which has a smaller variability, i.e., higher precision. In practice, there are other factors to consider that will affect these design choices and readers are referred to the previously mentioned publications for more detailed discussions on optimal and efficient designs.

4. Notes

1. As more complex experiments are designed, it is important to keep in mind how the gene expression studies fit into the overall research goal, and the limitations of gene expression studies. Consider, for example, that the long-term goal of an investigator is to understand the regulatory frame work behind the trans-differentiation between Type I and Type II cells in rat alveolar epithelium. If the investigator had simply performed a gene expression studies comparing these two cell types at a given point in time, the gene-expression analysis will be able to provide a list of differentially expressed genes at a given point in time. However, it will not be able to study or identify coordinated changes or co-expressed genes as there are only two conditions. A more thoughtful time-course experiment will be needed. Furthermore, more complex analysis involving integrating sequences and other biological meta-data will be needed to answer these types of questions.
2. The term *probe* refers to the DNA sequences that are spotted onto arrays. The cDNA or oligonucleotide probes are also called spots. Target refers to the samples that are hybridized to the arrays. Control samples are also called reference samples.
3. The purpose of normalization is to identify and remove systematic variation that occurs from the microarray experiment rather than real biological differences between the mRNA samples. The commonly used normalization for two-color microarray attempts to adjust for bias between the two dyes. It is often observed that the variation between red and green channels is not always constant i.e. bias can occur as a function of the intensity of the signal. Intensity-dependant variation can be corrected by generating a best-fit curve known as the "loess" line through the middle of an MA-plot (**Fig. 1**), and this becomes the new zero line for the vertical axis. In addition, intensity-dependent variation, spatial or print-tip bias could be a significant source of variation. Fitting a series of best-fit curves to different spatial or print-tip regions of the microarray is one method for adjusting spatial and intensity bias simultaneously. This commonly used method is known as the "print-tip loess" normalization.
4. Many normalization methods, such as global median and print-tip loess, rely on the assumption that majority of genes are not differentially expressed. In circumstances where there are large numbers of differentially expressed genes, normalization control probes are needed to be incorporated into the procedure to

ensure that the adjusted observed intensity indeed reflects the differential gene expression and not artefictual bias from the microarray experiment. Typically, housekeeping genes with a constant level of gene expression are the controls of choice in conventional quantitative PCR-based studies. However, these controls are not ideal for microarray studies because there may be tissue differences in the level of expression, and the housekeeping gene expression is at the upper range (and so not necessarily good controls for low expressing genes in the microarray). Thus, housekeeping genes are not representative of all intensity values expected in the microarray. To get around this problem a set of controls called microarray sample pool (MSP) has been designed (6). This control comprises the genomic DNA (minus intervening sequences) for all genes present in the microarray. DNA species making up the MSP are pooled and titrated at different concentrations. Thus, the MSP is potentially a control for all labeled cDNA sequences. As it is titrated at different concentrations, the MSP titration series will also allow for intensity-dependent normalization.

References

1. Wang, X., and Seed, B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* **19**, 796–802.
2. Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., De Risi, J. L., et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol.* **4**, R9.
3. Rouillard, J. M., Zuker, M., and Gulari, E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.* **31**, 3057–3062.
4. Tsai, J., Sultana, R., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., et al. (2001) RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol.* **2**, software0002.0001–software0002.0004.
5. Verdugo, R. A., and Medrano, J. F. (2006) Comparison of gene coverage of mouse oligonucleotide microarray platforms. *BMC Genomics* **7**, 58.
6. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M, Peng, V., Ngai, J., et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15.
7. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71.
8. Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., et al. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.* **33**, D562–D566.

9. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210.
10. Choi, J. K., Choi, J. Y., Kim, D. G., Choi, D. W., Kim, B. Y., Lee, K. H., et al. (2004) Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett.* **565**, 93–100.
11. Ghosh, D., Barette, T. R., Rhodes, D., and Chinnaiyan, A. M. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics* **3**, 180–188.
12. Cox, D. R. (1958) *Planning of Experiments*. Wiley, New York.
13. Cox, D. R., and Reid, N. (2000) *The Theory of the Design of Experiments*. Chapman and Hall, Boca Raton, FL.
14. Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978) *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York.
15. Cobb, G. W. (1998) *Introduction to Design and Analysis of Experiments*. Springer, New York.
16. Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U S A* **97**, 9834–9839.
17. Rodriguez, M. W., Paquet, A. C., Yang, Y. H., and Erle, D. J. (2004) Differential gene expression by integrin beta 7+ and beta 7– memory T helper cells. *BMC Immunol.* **5**, 13.
18. Freedman, D., Pisani, R., and Purves, R. (1988) *Statistics*. Norton, New York.
19. Page, G. P., Edwards, J. W., Gadbury, G. L., Yelisetti, P., Wang, J., Trivedi, P., et al. (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics* **7**, 84.
20. Seo, J., Gordish-Dressman, H., and Hoffman, E. P. (2006) An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics* **22**, 808–814.
21. Tibshirani, R. (2006) A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* **7**, 106.
22. Bolstad, B. M., Collin, F., Simpson, K. M., Irizarry, R. A., and Speed, T. P. (2004) Experimental design and low-level analysis of microarray data. *Int. Rev. Neurobiol.* **60**, 25–58.
23. Kendziorski, C. M., Zhang, Y., Lan, H., and Attie, A. D. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics* **4**, 465–477.
24. Kendziorski, C., Irizarry, R. A., Chen, K. S., Haag, J. D., and Gould, M. N. (2005) On the utility of pooling biological samples in microarray experiments. *Proc. Natl. Acad. Sci.* U S A **102**, 4252–4257.

25. Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Sabet, H., Tran, T., et al. (1999) The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb. Symp. Quant. Biol.* **64**, 71–78.
26. Kerr, M. K. (2003) Experimental design to make the most of microarray studies. *Methods Mol. Biol.* **224**, 137–147.
27. Yang, Y. H., and Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **3**, 579–588.
28. Speed, T., and Yang, Y. H. (2002) Direct versus indirect designs for cDNA microarray experiments. *Sankhya Ser A*, Part 3, **64**, 706–720.
29. Glonek, G. F., and Solomon, P. J. (2004) Factorial and time course designs for cDNA microarray experiments. *Biostatistics* **5**, 89–111.

6

Clinical Uses of Microarrays in Cancer Research

Carl Virtanen and James Woodgett

Summary

Perturbations in genes play a key role in the pathogenesis of cancer. Microarray-based technology is an ideal way in which to study the effects and interactions of multiple genes in cancer. There are many technologic challenges in running a microarray study, including annotation of genes likely to be involved, designing the appropriate experiment, and ensuring adequate quality assurance steps are implemented. Once data are normalized, they need to be analyzed; and for this, there are numerous software packages and approaches.

Key Words: cancer, microarray, annotation, experimental design, quality metrics, normalization, filtering, analysis.

Abbreviations: GO – gene ontology; RT-PCR – reverse transcriptase PCR

1. Introduction

Cancer is a genetic disease. Though this seems hardly worth mentioning nowadays, it is, in fact, a relatively new idea. Until the 19th century, cancer was believed (due to Hippocrates) to be a disease caused by an excess of black bile. At the beginning of the 20th century, the developmental biologist Theodor Boveri first suggested a linkage between tumor formation and improper chromosome segregation based on observations he made in sea urchins *(1)*. Since then, the discovery of the structure of DNA in the 1950s *(2)*, combined with genetic evidence of inheritance of cancer and the discoveries of oncogenes [via the Rous sarcoma virus and the src gene in the 1970's *(3)*] and tumor

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

suppressor genes, e.g., retinoblastoma in the early 1980s ***(4)***, led the way to a more general understanding of the causal processes involved in neoplasia.

That cancers are caused by changes at the DNA level is now a given. However, few cancers, especially those of epithelial origin (so-called solid tumors), make it to an advanced stage by the mutation of a single gene. Instead, current thinking is that tumorigenesis is a haphazard, stepwise process involving multiple short circuits of regulatory pathways, adaptation, avoidance of immune surveillance, and finally emergence of properties that allow disseminated growth—metastasis—typically the ultimate cause of death ***(5)***. Indeed, most cancers initiate with a single mutation, conferring a growth advantage compared to neighboring cells ***(6)***. The act of clonal expansion becomes associated with further mutations, which endow subsequent cell populations with invasive and metastatic potential. The path is often convoluted and is also associated with cul-de-sacs and cell death. Unfortunately, the selective pressures and numbers of events that accrue over the many years of tumor development are statistically high enough that cancer is a leading cause of death. While many of the key contributors and cellular pathways that push a cell into neoplasia have been determined, there are clearly many more that are unknown.

Identification of all the factors involved in cancer should provide not only greater understanding of the biology of this collection of diseases, but also new therapeutic targets and diagnostic markers. A critical milestone on the road to this understanding (and this is true for many other diseases) was the completion of the rough and final drafts of the human genome ***(7)***. Indeed, one of the most important findings by the genome sequencing project was that the actual number of genes that are present in humans, roughly 25,000 by last estimate ***(8)***, is far less than the previously anticipated number of ~100,000 ***(9)***.

With ~25,000 defined genes in the human genome, there is now a hard limit on the space through which biological causation for neoplasia has to be searched. This is not to say the task will not be formidable, since the relative paucity of human genes (compared with plants, for example) is countered by the complex processing of these genes through differential splicing, post-translational modifications, and exquisite control of function. There are probably at least 10-fold the number of human proteins compared to genes due to splice variants. For some proteins, there is already a plethora of knowledge as to their biological effect, either directly determined or gathered using a whole organism or systems biology approach. Unfortunately, for most genes and proteins, very little information exists regarding their biological activity. Also, each of the levels of information about a gene enriches our knowledge base but adds a level of complexity to any analysis, promoting informatics to the forefront if there is to be any hope of interpreting this deluge of data. Integrating all of this information in a cohesive and meaningful way is one of the goals of bioinformatics ***(10)***. The Cancer Genome Anatomy Project, the Human Cancer Genome Project, and the Cancer Genome Project represent three large-scale programs attempting to coalesce a wide range of cancer related genomic data into a single resource ***(11)*** (*see* **Note 1**).

2. Microarray Technology

Microarrays leverage biological information with bioinformatics knowledge to bring new insights to many biological systems. They have been applied most intensively to the field of cancer research. The basic premise behind a microarray is that thousands of fragments of DNA (the probes) representing various genes are attached to the surface of an inert material. Reverse transcribed messenger RNA labeled with a fluorescent dye from an experimental sample (the target) is then co-hybridized on the microarray. After removal of non-selectively bound fluorescent material, the microarray is scanned at a high resolution to quantify the amount of fluorescent signal over the surface. Since the locations of the gene probes are pre-determined, the relative or absolute amount of RNA for each of the genes on the array can be calculated. A microarray therefore measures the levels of mRNA transcripts in a sample. Since thousands of gene fragments can be located on an array, it can provide a genome-wide view of gene expression in cancer.

The two most popular microarray technologies employ the use of either double-stranded cDNA probes spotted onto slide surfaces, pioneered by the work of Pat Brown's laboratory ***(12)***, or single stranded oligonucleotide probes. Oligonucleotide microarrays are made either by depositing presynthesized oligonucleotides directly onto a slide surface mechanically, e.g., using ink jet technology (www.agilent.com), or assembled base by base using photolithography in a process somewhat analogous to semiconductor etching using photoactivatable nucleotides and multiple masks. The photolithographic technique was developed by Stephen Fodor in the early 1990s ***(13,14)*** and is the basis behind the microarray platform used in the Affymetrix GeneChips (www.affymetrix.com).

A comparative approach is typically taken to identify genes important in neoplasia. For example, tumorous tissue may be compared to normal tissue, or tumors of different pathological types or grades may be looked at (for an example of the latter *see* **Chapter 8**). The relative fold difference between the categories is determined for each particular gene being measured on the microarrays, and significant groups of genes can be aggregated based on similar behavior and looked at either as a whole or individually. Morphologically indistinct tumors may actually belong to clinically and biologically distinctive categories. For example, Her2/neu/ERBB2 positive status in breast cancer is widely used to assess trastuzumab (Herceptin) treatment, which is, in fact, the target of the drug ***(15)***. Microarrays are now routinely used in the hope of finding such biomarkers.

Microarray expression profiling of cancers has now been used in a wide variety of settings. Correlation to pathological subtype was first shown by Golub et al. ***(16)*** in comparing acute myelogenous leukemia and acute lymphocytic leukemia. In malignant lymphomas, Alizadeh et al. ***(17)*** were the first group to show a correlation to therapeutic outcome with gene expression profiles in B-cell lymphoma. A distinct signature for metastatic tumors was found by Ramaswamy et al. ***(18)*** in primary lung adenocarcinomas. Meta-analysis of gene expression datasets has been used to find a generalized signature for proliferation ***(19)***. Prediction of patient outcome was demonstrated successfully in breast cancer by grouping genes together into "meta" genes ***(20)*** (*see* **Note 2**). A commercially available diagnostic based on microarray data for the recurrence of breast cancer in tamoxifen-treated patients is now available ***(21)***. Extending the concept of finding gene signatures, grouping genes according to the pathways they are involved in has also proved successful ***(22)***. In a recent study, Glinsky et al ***(23)*** combined data from mouse and human prostate tumor samples to derive an 11-gene signature that consistently holds a stem-cell type expression profile predicting poor outcome across a variety of cancer types. Microarrays have also been used extensively to profile cancers of the prostate ***(24,25)***, ovaries ***(26)***, kidney ***(27)***, breast ***(28,29)***, and lung ***(30,31)***, along with a number of other neoplasms ***(32–34)***.

For the remainder of this chapter, we discuss the various factors involved in the design and interpretation of a typical microarray project focused on a clinical application. **Figure 1** illustrates a general flow chart of the process. The approach will necessarily be a general one, and the merits of various options at each stage will be discussed. We will focus on techniques specific to cDNA arrays, though comparable methodologies exist for oligonucleotide arrays. The "Tumor Analysis Best Practices Working Group" has published an extensive review of microarray techniques focused on Affymetrix arrays ***(35)***.

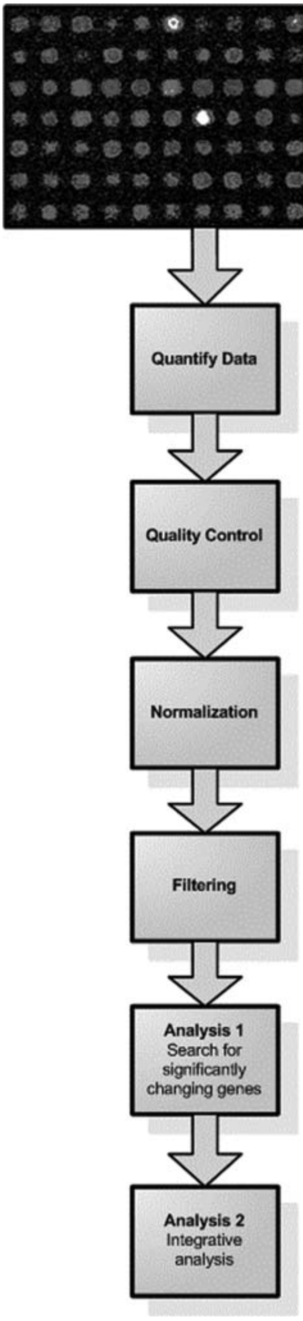


Fig. 1. Generalized process flowchart for analyzing microarray data.

3. Design and Interpretation

3.1. Annotation

One of the most important informatic steps that needs to be done before a single microarray has even been hybridized is annotation of the array platform. Out-of-date annotation can severely impact interpretation of the dataset, as Dai et al. *(36)* have shown. The basic information as to what a spot on an array is, either EST accession ID or sequence information, is the first starting point. However, this information is somewhat meaningless unless placed in an appropriate context with higher levels of meta-annotation. Unfortunately, all of the meta-annotations one might need to use are in disparate public and private databases. These databases are themselves dynamic in nature and therefore constantly changing at different rates, making contemporary annotation a non-trivial task *(10)*. The SOURCE database attempts to pull together various annotations for users of microarrays, and is also updated on a regular basis *(37)*; however, the information it contains is only basic. As a consequence many producers of microarrays now invest substantial effort in maintaining array annotation. At the University Health Network Microarray Centre in Toronto, for example, all databases and annotations on arrays are fully re-annotated quarterly and this seems to be a reasonable schedule.

A database such as mySQL (www.mysql.org), installed on a server running Linux (http://distrocenter.linux.com/platforms), is a good starting point for pulling data annotations together. There is no "one size fits all" approach to annotation, and much will depend on a researchers' particular interest. The starting point usually involves retrieving sequence for each of the elements on an array and finding their genomic position using BLAT *(38)*. For cDNA probes, searching for repeats in sequences using REPEATMASKER *(39)* is important, since these can confound results if not taken into account. This should not be an issue for oligonucleotide arrays as the design of the representative oligonucleotide includes avoidance of non-unique sequences.

Once a gene's position in the genome is known, further annotation can usually be derived by searching through the UCSC "knowngene" data table, available as a free download (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz). From this table a reference to kgXref (known gene cross-references table, also available at UCSC) can be used to find UNIPROT *(40)* and NCBI REFSEQ protein ids *(41)*. With the REFSEQ id, a link can be made to the UNIGENE *(42)* database, available for free by downloading and searching directly (ftp://ftp.ncbi.nih.gov/repository/UniGene). Unigene is a database of non-redundant gene ids assembled by clustering

GenBank sequences. Although commonly used, it can be unreliable ***(43)*** and it is therefore recommended that links to ENTREZ Gene ids ***(44)*** be made. This can be done by downloading and searching the table gene2accession found at NCBI (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2accession.gz). With ENTREZ Gene ids, links to OMIM can be made through the table mim2gene (ftp://ftp.ncbi.nih.gov/gene/DATA/mim2gene.gz). OMIM ***(45)*** is a curated catalog of human genes related to genetic disorders. Linking genes to actual published data found in PubMed ***(42)*** can be made using the gene2pubmed table (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz). Finally, Gene Ontology ***(46)*** annotations can be derived using the gene2go table (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz). Gene Ontology (GO) classifications are quite useful in pulling together higher level concepts surrounding a gene product. For example, a gene product may be part of the parent GO term "signal transduction." Further annotation is limited only by how deep a researcher wants to go, as many hundreds of specialized databases exist. The journal *Nucleic Acids Research* (http://nar.oxfordjournals.org) publishes yearly an issue devoted to the most popular online molecular biology related databases (*see* also **Chapter 17** for a summary of web-based resources of relevance).

3.2. Experimental Design

3.2.1. General Considerations

Suffice to say the experimental design is the most important stage for any scientist wishing to use microarrays in their clinical research (*see* **Chapter 5** for more details on this aspect of a microarray experiment). A good experimental design relies on a good hypothesis and appropriate scaling of the power of the experiment to the type and availability of the samples to be profiled. For example, one hypothesis might be that there is a difference in gene expression between normal and tumor samples from the same group of patients. Or, that there is a difference between two types of cancer present in the same tissue types in different patients (lung squamous cell carcinoma versus lung adenocarcinoma). The experimental conditions should then be set up according to the general statistical principles which all experiments are guided by. Care must be taken to balance the sample population for extraneous factors: smokers and non-smokers, age, sex, tumor stage, and so on. These types of factors may all effect gene expression and an obvious bias in one of the experiment's hypothesis driven categories will be confounded if not balanced correctly. Furthermore, it is now well known that variation between studies of the same type of sample are influenced by the samples themselves and by human related factors ***(47–49)***. Therefore, one must balance between experimental groups the technicians responsible for processing and handling samples, the slide lots used during the experiment, and possibly even the days during which RNA extraction and hybridization are being performed.

In two-color experiments in which two RNA samples are labeled with two fluorescent dyes with distinct emission spectra, one channel is usually designated for the experimental channel and one for a suitable control channel. There are two generally accepted methodologies for choice of control channel. In the first, the control channel is not a control per se, but is instead one of the experimental samples from one of the classes being examined, e.g., tumor type A versus tumor type B. In the "balanced block design" ***(50)*** each sample from the two classes is matched up to a member of the other class and the two are hybridized on the same array. In the "loop design" ***(51)***, each sample is split into two sub-samples and each sub-sample is used to connect arrays together in a loop pattern. In the other methodology, a reference sample is used as a control on each chip and comparisons between classes is accomplished by comparing ratios of experimental to control. This is by far the most common design for two-color cDNA experiments. It has the disadvantage that more microarray chips are needed compared to a block design since each sample has to be run on a separate chip. But one advantage is that multiple experiments (such as future studies that are done as follow up) using the same reference sample can be compared directly. Also, in cases where the object is to find novel classes based on expression patterns, only the reference design can be used. Finally, a reference design reduces the need to perform dye-flips ***(52)*** to some extent, because the reference samples are all labeled with the same dye between which comparisons are being made. Dye-specific bias is also less of an issue when using indirect labeling ***(53)***.

It is worth mentioning the value of profiling normal tissue types when deciding on an experimental setup. It is quite easy to find groups of genes which vary across normal tissue types ***(31,54)***. This is likely due to the heterogeneous nature of pathological samples, where a bias in tissue type, cells at a particular phase of the cell cycle or hormonal status, or even inter-individual expression signatures may cause certain genes to change between samples. Obviously it is of value to remove such genes from further analysis during the filtering stage.

3.2.2. Sample Numbers

Once a general experimental design has been decided upon, the next important choice will be the number of samples to be profiled. Unfortunately, there is no easy answer for determining sample size. The multiplex nature of microarray studies, many thousands of genes versus small numbers of samples, makes power calculations that are traditionally used in setting up clinical trials difficult to apply, and often dependent on the final analysis regimen. A point worth mentioning is that choosing a sample size based on traditional estimates of statistical power is fine when all that is necessary it to make an estimate on the confidence of the error term, but this can yield different numbers than what is necessary for assessing the power of a classification algorithm. Unfortunately, all of this is compounded by the fact that final analysis of microarray data is usually more exploratory and unknown at the start of a study. More often than not, the choice of sample size will be governed by the amount of pathological samples that are actually available, or the cost limitations. Still, there are some methods available for determining appropriate sample size ***(50,55–57)***.

Tibshirani's method ***(56)*** of sample size estimation is a good choice and does not assume equal variances or independence of genes on an array. The fact that it is based on estimating false discovery rates, a common methodology for finding genes that vary between groups, makes it even more suitable for use. The basic premise behind Tibshirani's method is to start with the output from a typical permutation based analysis routine and then estimate false discovery and false negative rates for different sample sizes. By way of example, to find a two-fold difference between two groups of 10 samples, e.g., normal versus tumor, each on an array with 1,000 genes, the sample size must be increased to roughly 30 samples in each group to get optimal results.

3.3. Quality Metrics

Measuring the quality of slides after hybridization and scanning is an important part of analysis. Visually inspecting images for gross defects should always be done. However, more robust methods are recommended. The best software for assessing data quality is found in the LIMMA package which is part of the Bioconductor ***(58)*** framework for the R (www.R-project.org) statistical language. We currently use Bioconductor version 1.8 with R version 2.3.1. Although there is not one particular number or function which will definitively point to a chip being of "pass" or "fail" quality, the joint use of these functions allows for an overall assessment to be made and for outliers, i.e., bad slides, to be identified.

Listing 1 shows a typical R script that can be used to assess raw data files after quantification. The script will output a series of quality related graphs into the working directory. The data files in this case were all quantified in ArrayVision (www.imagingresearch.com/products/ARV.asp).

Listing 1: An R script for assessing the quality of microarray slides in an experiment.

```
Library(limma)
files<-dir(pattern="^[1-9].*.txt”)
RG<-read.maimages(files,source=“arrayvision”)
RG$genes<-readGAL()
RG$printer<-getLayout(RG$genes)
spottypes<-readSpotTypes()
RG$genes$Status<-controlStatus(spottypes,RG)
MA<-normalizeWithinArrays(RG,method=“none”,
bc.method=“none”))
fnc<-file.path(“.”,paste(“F-RGdensities”,”.png”,sep=”))
png(filename = fnc, width = 6.5 * 140, height = 10 *
140,pointsize=20)
plotDensities(MA)
imageplot3by2(RG,z=“R”,low=“white”, high=“red”)
imageplot3by2(RG,z=“Rb”,low=“white”, high=“red”)
imageplot3by2(RG,z=“G”,low=“white”, high=“green”)
imageplot3by2(RG,z=“Gb”,low=“white”, high=“green”)
fnc<-file.path(“.”,paste(“F-boxplot”,”.png”,sep=”))
png(filename = fnc, width = 6.5 * 140, height = 10 *
140,pointsize=20)
par(ps=8,las=2)
boxplot(MA$M~col(MA$M),names=colnames(MA$M),
main=“boxplot”)
numarrays <- ncol(RG)
numpages <- ceiling(narrays/6)
for (ipage in 1:numpages)
{
i1 <- ipage * 6 – 5
i2 <- min(ipage * 6, numarrays)
fnc <- file.path(“.”, paste(“MA plot”, “-”, i1, “-”, i2,
“.png”, sep = “))
png(filename = fnc, width = 6.5 * 140, height = 10 *
140,pointsize = 20)
par(mfrow = c(3,2))
for (i in i1:i2)
{
plotMA(MA[, i])
}
}
```

Line 1 Loads the appropriate LIMMA module from Bioconductor into R.

Lines 2–7 Load the data (files ending in ".txt" and starting with a number from 1 to 9) in the current working directory and associate the correct layout of the array with the data itself.

Line 8 Copies the data into a new matrix called "MA" and can be used to change normalization method (method="none" in this case) or background subtraction choice (bc.method="none"). The choice of performing background subtraction is not always obvious. The underlying hypothesis of background subtraction is that the intensity of a spot on a microarray is the combined value of the background intensity around the DNA spot plus the foreground measurement of intensity within the limits of the actual spot. This hypothesis is somewhat flawed though, as negative control spots, i.e., water or DMSO spots with no DNA in them, often are of lower intensity than the surrounding background (an effect termed "ghosting"). As such, background subtraction can often introduce rather serious artifacts in the data. However, it can be useful if the slide surface consistencies are non-uniform, e.g., if fluid was not distributed evenly during hybridization. If considering background subtraction, it is recommended that the procedures in this section be repeated with background subtraction applied to see what effect it may have on the data.

Lines 9,10 Are used to set a filename and size of graph that will serve as the output from the next line.

Line 11 Shows a density histogram is generated. **Figure 2A** shows a typical density histogram for a series of 20 slides. The red and green lines both overlap substantially and the curves are roughly symmetrical around the peaks. Compare this to a series of slides (**Fig. 2B**) where there was a problem during RNA extraction and labeling. The histograms are very inconsistent and a problem is immediately assessed. In this case the entire experiment should be repeated.

Lines 12–15 Plot pseudo-colored spatial images of the data for both channels. Separate files for both channels and the foreground (spot) and background intensities are generated. **Figures 3A** and **3B** show a typical output for the green channel foreground and background signal of two slides with acceptable quality.

Notice that there is a lot of consistency across the slide surface. Contrast these two images with **Figures 3C** and **3D**, which obviously display a problem that occurred during hybridization. These slides will probably need to be repeated.

Lines 16–19 Setup and output a boxplot of all the current slides into a file. A boxplot shows the overall mean (horizontal line in the centre of the box) along with the interquartile range (the box) and outliers (circles). Once again, relative consistency is appropriate, and a box plot can

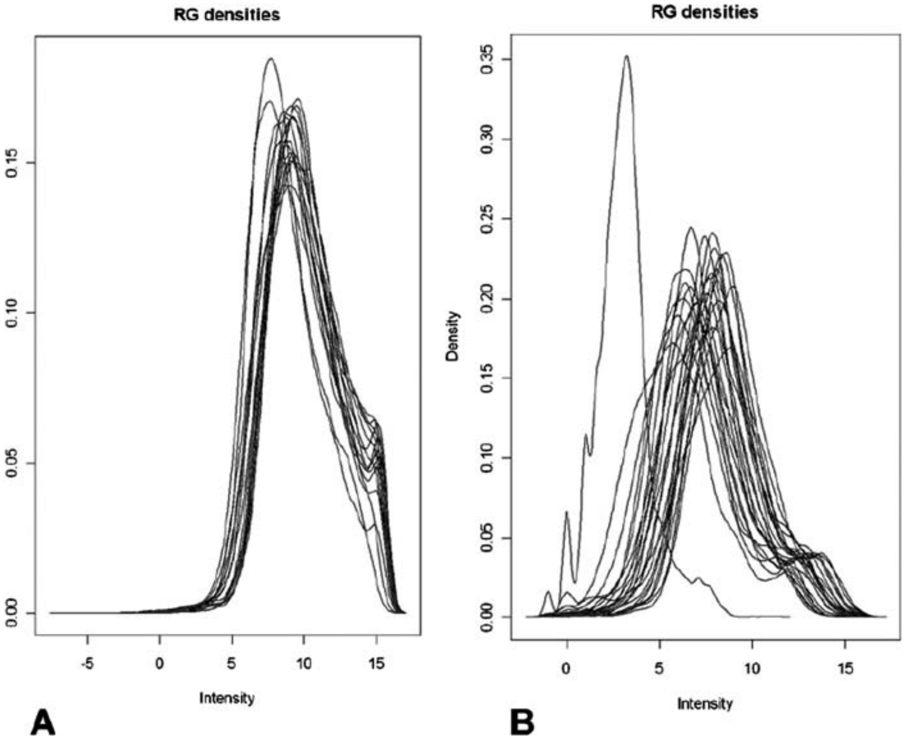


Fig. 2. **A:** A typical density histogram for a series of 20 slides. The lines represent density measurements from individual slides on both channels. **B:** A series of slides where there was a problem during RNA extraction and labeling.

allow for an outlier slide to be easily determined. In **Fig. 4** it is easy to see that slide number 5 is different from the other slides, and closer examination of this slide would be necessary to determine if it should be accepted or rejected for the study.

Lines 20–33 Iterate through all the arrays loaded into the dataset and output MA plots (6 plots per page as determined on line 28). An MA plot shows any intensity dependant effects in the raw data where M = $\log_2$(R/G) and A = (0.5)*$\log_2$(R*G). The x-axis represents a log ratio of equal intensities in both channels. The majority of data in a typical microarray dataset should spread out roughly along this line. **Figure 5A** shows an MA plot from a slide with a good distribution of data. **Figure 5B**, on the other hand, clearly has an intensity-dependant issue that goes beyond that which can be corrected later by normalization.

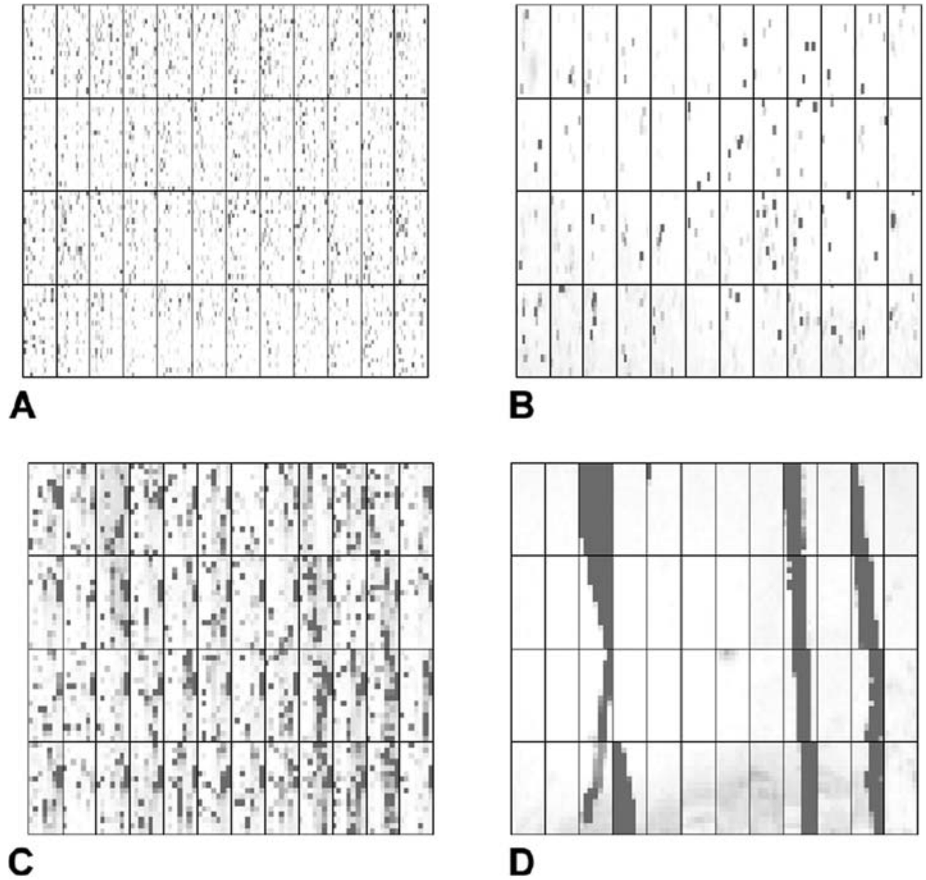


Fig. 3. Foreground **(A)** and background **(B)** signal intensities for a slide with a suitable degree of quality. Foreground **(C)** and background **(D)** signal intensities for a slide with a problem that occurred during hybridization.

3.4. Normalization and Filtering

3.4.1. Normalizing Data

Once assured of some measure of slide quality, the data will first need to be normalized before valid comparisons between sample subjects are made. This is because differences in labeling efficiencies, amounts of starting RNA materials, fluctuating hybridization conditions and other biases may be present from slide to slide. The box plots in **Fig. 4** (aside from the problem slide) show that the

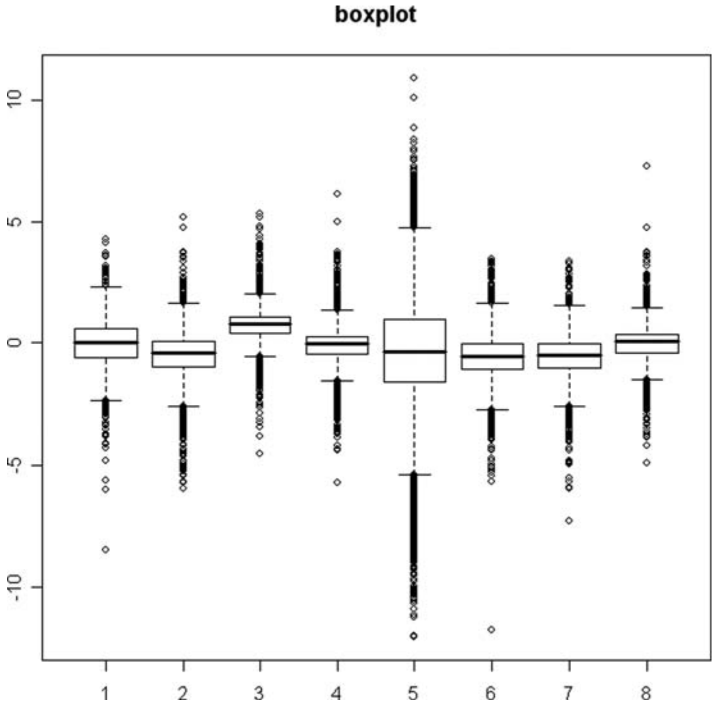


Fig. 4. Boxplot diagram for eight slides with varying degrees of scatter and quality. A boxplot displays summary statistics such as central tendency and variability for each slide.

distribution for individual slides is not quite the same, even though they are expected to be. This platform related variation should be accounted for before looking for biological variation. The most commonly applied normalization is a “per-chip” normalization that balances out systematic biases within a slide itself. It is well known that there are intensity-dependant issues in microarray ratio measurements (with low intensity measured log ratios typically deviating from zero) and that these furthermore vary across pin groups on spotted slides ***(59,60)***. Therefore, typically a locally weighted linear regression curve (loess) is fit on a pin-to-pin basis to an MA plot of the data and values in the control channel are adjusted up or down to correct for deviation from this line. **Figure 6A** shows an MA plot of a slide before loess normalization and **Figure 6B** shows the data after normalization. Further normalization usually consists of normalizing each genes’ log ratio to zero across chips to account for genes that are either highly or lowly expressed across all samples. It is

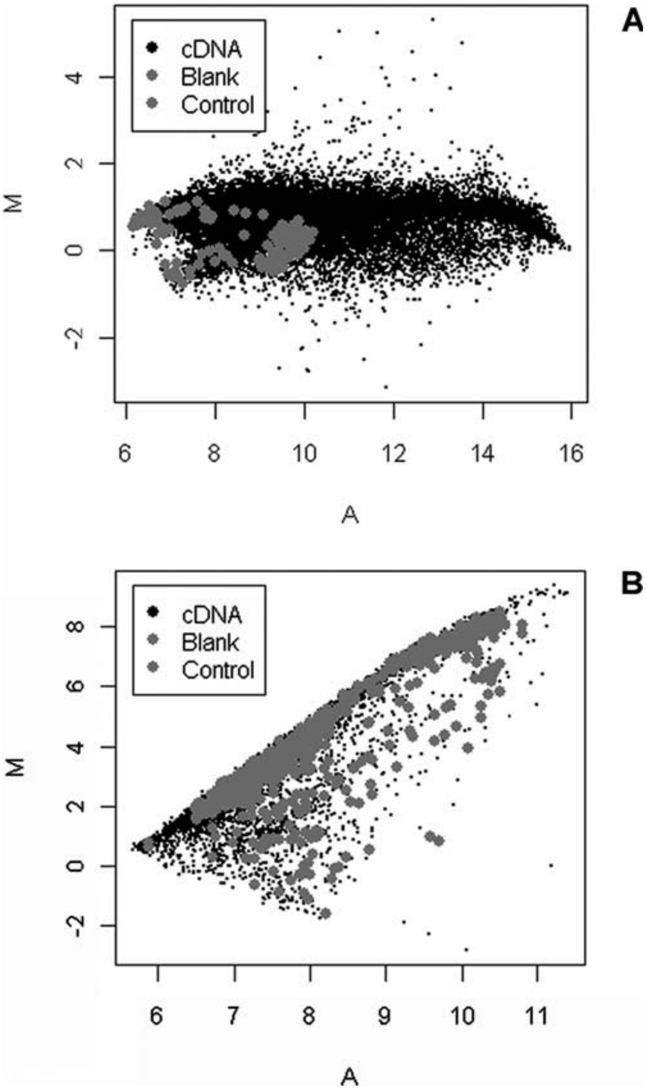


Fig. 5. M (log intensity ratios) versus A (log total intensity) plots showing scatter from a typical slide of **(A)** good quality and **(B)** poor quality. An MA (intensity scatter) plot compares intensity on two colors (or two chips).

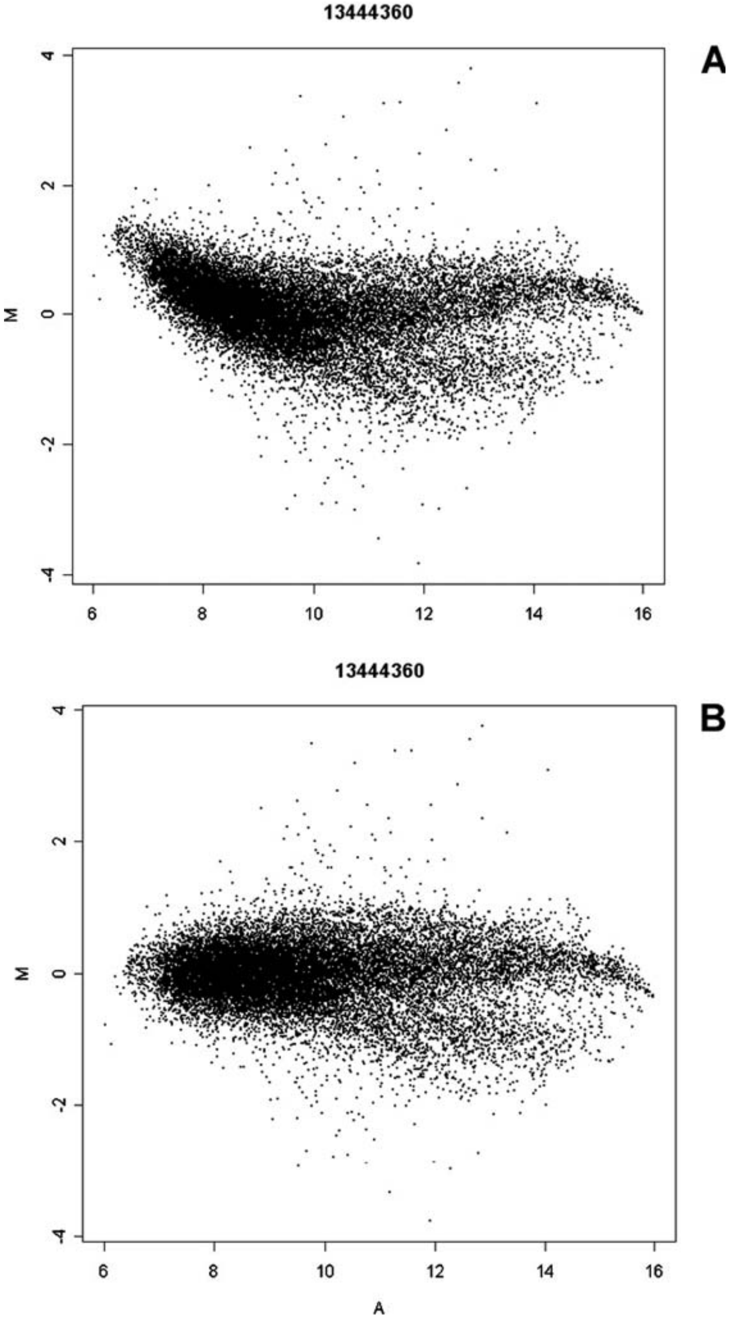


Fig. 6. MA plot of a slide **(A)** before loess normalization and **(B)** after loess normalization.

usually easier to visualize genes that are changing between samples after such a normalization. Normalization is a standard procedure and can be performed in any microarray data analysis package. For example, to loess normalize data in R, simply change line 8 in Listing 1 to:MA<-normalizeWithinArrays(RG, method="loess", bc.method="none")).

3.4.2. Filtering Unwanted Genes

Filtering is usually the next step in microarray data analysis. Filtering is an important step because it removes genes from further examination that are not of interest. Genes with very low intensity are typically removed because there tends to be much higher variability with their measurements. A good rule of thumb is to flag or remove those genes with intensities that are less than two times the average background intensity before normalization. Ranking genes in percentiles is another way to determine an appropriate cutoff. For example one might wish to remove those genes in the lower 10th percentile of overall intensity measure. Filtering of genes can also be done when there is *a priori* information of spots which may be problematic. On spotted cDNA slides, there are sometimes problems related to production e.g. during amplification of spotted material, which may warrant some spots being removed. It is recommended to remove spots which have a high percentage of repetitive elements in them. Repeat elements such as the human ALU repeat are present in upwards of 10% of the human genome ***(61)***, and these are frequently present in the non-coding ends of spotted cDNA's on microarrays. Since we would expect an inordinate amount of cross hybridization in sequences containing a large amount of repeat (say, greater than 33%) they should be filtered out. Finally, as mentioned earlier, there are many genes which vary naturally to a great extent between individuals or samples. If a study is done with an appropriate number of normal samples, one can first look for genes that vary between individuals and filter them out from subsequent analysis when looking at differences between tumor types.

3.5. Analysis

3.5.1. Differentially Expressed Genes

Analyzing microarray data usually first takes the form of finding genes which are differentially expressed, either between experimental and control channels on chips or between samples. Finding differentially expressed genes is usually done first, both to find genes of interest and to further filter data before application of more sophisticated data mining techniques such as clustering.

When attempting to find genes that are over or under expressed one typically chooses a threshold, such as a twofold difference. This number was originally determined by concordance analysis for one dataset ***(62)*** but has become a guideline criteria now used in many different analyses. More sophisticated measures, such as using a Z-score ***(63)*** to estimate fold changes in an intensity dependant manner can also be used. A one-sided t-test can be performed if replicates were done for each sample. This must be multiple-test corrected (*see* next section 3.5.3). Once a threshold is decided on, the usual course is to apply that threshold to being found across a percentage of the samples. For example, if 100 samples were obtained and profiled, one might choose to look only at genes that have at least a twofold difference in 33% of them. Choosing a twofold change level will undoubtedly lead to removal of true differences that are lower, but it will still allow for finding the less conservative changers. This is a testament to microarrays being a screening technology where one is usually looking for the "low hanging fruit."

3.5.2. Categorizing Samples

Attempting to categorize samples can be done in one of two (or both) ways. Unsupervised methods are exploratory in nature. Agglomerative hierarchical clustering is one such technique. In this method, two genes that have the most similar expression profiles across experiments, based on a similarity measure such as their Pearson Correlation, are found. The average is taken between these two genes and then a new gene most similar to this "average gene" is found in the rest of the set. The process is iterated and a tree type diagram can be built up (**Fig. 7**). The length of tree branches is related to the degree of similarity between adjoining groups. Individuals sample are similarly clustered according to their nearest neighbors. A cluster diagram allows one to explore those categories of genes or samples which are nearest to one another and hypothesis regarding biological meaning can be generated. For example, it is often hypothesized that genes that are closely clustered together are related to one another in some manner, such as belonging to the same molecular pathway. Care should be exercised in interpreting clusters since clustered genes can rotate around a branch and hence distance between genes on the edges of neighboring clusters is much larger than genes within a cluster. Cluster diagrams were first applied by Eisen ***(64)*** to microarray data and are now very common. Some clustering methods, such as those with bootstrapping, aim to assess the statistical significance of tree branch locations ***(65)***. Principal components analysis is another unsupervised method to attempt to find biologically meaningful groups in microarray results ***(66)***. By projecting the multiple dimensioned space of

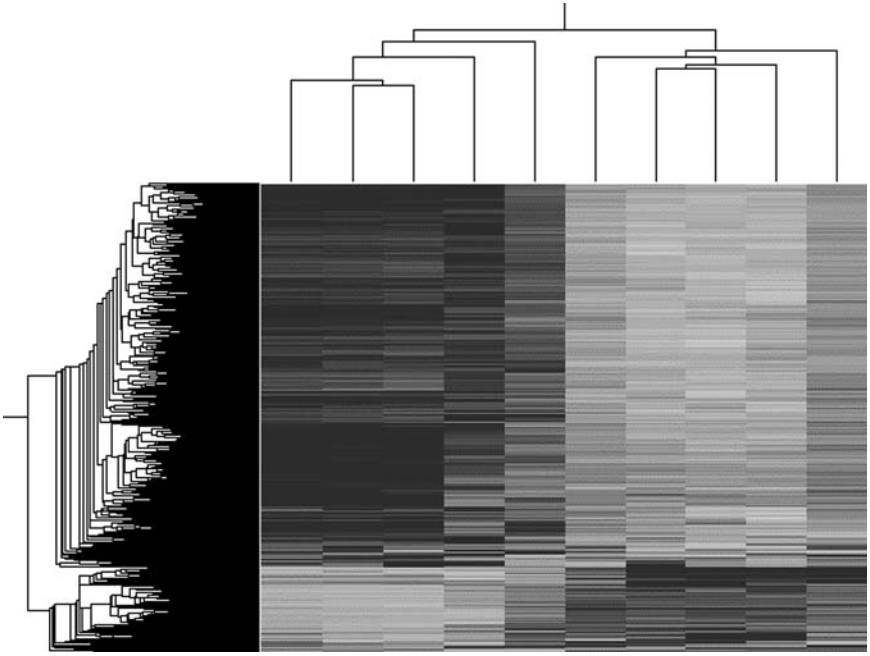


Fig. 7. A typical dendogram obtained after performing a two-way hierarchical clustering Branches on the left indicate genes which cluster together according to similarity of expression. Likewise, branches across the top indicate the degree of similarity of individual samples. In this case, it is evident that there are two distinct groupings of samples according to their expression across a wide number of genes.

a microarray dataset onto axis in lower dimensional space, those samples or genes most closely related will be grouped together (*see* **Chapter 2** for more discussion on principal component analysis).

3.5.3. Data Mining

Supervised methods of data mining are used when a priori information regarding categories in the data exist. The most common method is simply to perform a t-test or an ANOVA ***(67)***. However, when performing such tests it is necessary to correct for multiple testing of hypothesis. With very many genes on an array, performing multiple t-tests or ANOVA's will result in many false positive results. For example, at the 5% significance level for an array with 10,000 genes, one would expect 500 genes to be falsely identified as varying. Therefore, a correction to the test statistics must be made. The simplest is a

Bonferonni correction, where the significance *p*-value is simply weighted by dividing its value by the number of tests. This is usually too strict though, and typically something such as a Benjamini and Hochberg ***(68)*** false discovery rate multiple test correction is used.

A popular supervised mining technique is "significance analysis of microarrays" (SAM), first developed by Tibshirani ***(69)***. SAM is similar to a t-test but uses a term called a "d-value" instead. If there are two categories of samples, I and U, then let $\overline{x}_I$ and $\overline{x}_U$ be the mean expression measurements of gene *i* in the two categories, respectively. The test statistic is then:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_o}$$

where

$$s(i) = \left[\left(\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2}\right)\left\{\sum\nolimits_m [x_m(i) - \bar{x}_I(i)]^2 + \sum\nolimits_m [x_n(i) - \bar{x}_U(i)]^2\right\}\right]^2$$

n_1 and n_2 are the numbers of expression values in each category, and Σ_m andΣ_n are summations of the differences of expression measurements from the mean expression measurement in categories I and U, respectively. The idea is then to estimate for each gene an expected value of d by randomly shuffling across all samples the expression values, recalculating d and then taking the average value after a fixed number of iterations. When plotting the observed d-values versus the expected d-values, any gene that deviates far from the "observed=expected" line is likely to be significant. A tuning parameter can be chosen which sets the upper and lower bounds on how far a deviation has to be before being called significant.

Other popular supervised data mining techniques include "gene-shaving" ***(70)***, which can be also used in an unsupervised fashion, and "weighted-voting," which was first introduced by Golub ***(16)***. Of course, there are many more methods for mining data and it is often overwhelming trying to decide on a course of analysis. A good approach is to run multiple analyses and look for common pools of genes that are significant across all methods.

3.5.4. Validation

Validation is crucial at the end of the analysis pipeline. Although every attempt is made to control for extraneous factors affecting gene expression measurements, including statistical techniques to minimize false positive results, nevertheless there will be mistakes. Currently the best methodology to validate results is RT-PCR *(71)*. Microarrays have a low dynamic range of measured gene expression and tend to underestimate changes, whereas RT-PCR has a high dynamic range. Hence, validation is used to assess overall observed trends rather than duplicate results.

3.6. Higher-Level Analysis

With lists of significantly co-regulated genes in hand, the next step is to place them into a biologically meaningful context. Comparing GO annotations in lists of genes is a good starting place. The basic premise here is that there may be over-representations of particular traits in one list of genes compared to another and this may represent unique biological features. To do this, an appropriate statistical measure to compare gene lists must be used, such as a chi-square or Fisher's exact test. Of course, as with previous statistical tests, a correction for multiple hypothesis testing must be made, and methodologies for this exist *(72,73)*. One can also combine GO annotation with expression data during the data mining phase of analysis *(22)*.

It is well known that genomic instability, a distinctive trait of cancer, leads to increased or decreased copy numbers of certain genes. That this will have some effect on gene expression is probable; whether this effect is linear is unknown. However, by integrating expression data with array comparative genomic hybridization (CGH) data, this question can be at least partially answered. Bussey et al. *(74)* looked at copy number and expression in the NCI-60 panel of human cancer cell lines and found a positive correlation between the two. Somewhat related to this, integrating genomic information for a particular region with mRNA expression and protein data allowed Mootha et al. *(75)* to find a candidate gene for cytochrome c oxidase deficiency.

Reconstructing genetic networks is another area that has been well explored. The underlying hypothesis is that correlation in expression can be related to co-regulation and hence involvement in similar pathways. The representational analysis presented earlier using GO functional categories is one way to dissect out these interactions. Combining protein-protein interaction data with gene expression measurements is another method that has been used *(76)*. Expression data combined with phylogenetic conservation has also proven successful *(77)*. Indeed, there is some evidence that genes that are co-expressed are often clustered together on the genome, hence bringing issues of selective pressure into play *(78)*.

4. Conclusions

There is no single approach in using microarrays to study neoplasia. Instead, there are a number of methodologies that can be chosen depending on the questions being asked. Here, we have presented a general methodology for the design and analysis of experiments using microarrays. With careful selection and appropriate controls, new insight into the functional underpinnings of cancer can be studied leading to better diagnostics and possible future treatments. The most critical element in employing microarray technology is selection of the most appropriate experimental design that complements the sample set available. It is far simpler to alter analysis strategies of a good dataset than to try to remediate deficiencies in a poor one.

5. Notes

1. The three cancer-related programs described involve centers in the UK, USA, and Brazil ***(11)***. The extensive molecular analysis of various cancers from the programs has DNA sequencing as the common link. The output from this work involves the sequencing of cDNA libraries, SAGE libraries (SAGE—Serial Analysis of Gene Expression is another form of microarrays), and mutation testing of genomic DNA from various tumors. Without a planned and informatics-based approach the accumulated data would not be available to the wider research community, and without this integrated approach the effectiveness of data mining would be diminished.
2. Meta analysis is a well-accepted approach in evidence based medical practices and works through statistically combining the results of previous experiments or studies. In the analysis of microarray studies, it is proposed in ***(20)*** to optimize the data generated by combining through metagenes, i.e., the aggregate patterns of variation of subsets of potentially related genes. The example given is breast cancer. Predicting prognosis in this cancer is a major clinical challenge. Although staging through assessment of lymph node involvement, estrogen receptor status, age and tumor size are major prognostic factors, it still is not possible to predict consistently the outcomes. Using various statistical approaches a combination of clinical and genomic (metagene) data is likely to be used increasingly to guide decision making.

References

1. Balmain, A. (2001) Cancer genetics: from boveri and mendel to microarrays. *Nat. Rev. Cancer* **1**, 77–82.
2. Watson, J. D., and Crick, F. H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967.

3. Stehelin, D., Varmus, H. E., Bishop, J. M., and Vogt, P. K. (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173.
4. Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., et al. (1986) A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646.
5. Hanahan, D., and Weinberg, R. A. (2000) The Hallmarks of Cancer. *Cell* **100**, 57–70.
6. Nowell, P. C. (2002). Tumor progression: a brief historical perspective. *Semin. Cancer Biol.* **12**, 261–266.
7. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
8. International Human Genome Sequencing Consortium. (2004) Human genome sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
9. Pennisi, E. (2003) Bioinformatics: gene counters struggle to get the right answer. *Science* **301**, 1040–1041.
10. Stein, L. D. (2003) Integrating biological databases. *Nat. Rev. Genet.* **4**, 337–345.
11. Strausberg, R. L., Simpson, A. J. G., and Wooster, R. (2003) Sequence-based cancer genomics: progress, lessons and opportunities. *Nat. Rev. Genet.* **4**, 409–418.
12. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.
13. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773.
14. Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., and Adams, C. L. (1993) Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556.
15. Ludwig, J. A., and Weinstein, J. N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* **5**, 845–856.
16. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
17. Alizadeh, A. A. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.
18. Ramaswamy, S., Ross, K., Lander, E., and Golub, T. (2003) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54.
19. Rhodes, D. R., and Chinnaiyan, A. M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet.* **37**(Suppl), S31–S37.

20. Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., et al. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci. U S A* **101**, 8431–8436.
21. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826.
22. Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37**(Suppl), S38–S45.
23. Glinsky, G. V., Berezovska, O., and Glinskii, A. B. (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* **115**, 1503–1521.
24. Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., et al. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826.
25. Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., et al. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U S A* **101**, 811–816.
26. Schaner, M. E., Ross, D. T., Ciaravino, G., Sorlie, T., Troyanskaya, O., Diehn, M., et al. (2003) Gene expression patterns in ovarian carcinomas. *Mol. Biol. Cell* **14**, 4376–4386.
27. Vasselli, J. R., Shih, J. H., Iyengar, S. R., Maranchie, J., Riss, J., Worrell, R., et al. (2003) Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc. Natl. Acad. Sci. U S A* **100**, 6958–6963.
28. Sorlie, T., Perou, C., Brown, P., Botstein, D., and Borresen-Dale, A. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. U S A* **98**, 10869–10874.
29. Sotiriou, C. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U S A* **100**, 10393–10398.
30. Garber, M. E. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. U S A* **98**, 13784–13789.
31. Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., et al. (2004) Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* **363**, 775–781.
32. West, R. B., and van de Rijn, M. (2006) The role of microarray technologies in the study of soft tissue tumours. *Histopathology* **48**, 22–31.
33. Tinker, A. V., Boussioutas, A., and Bowtell, D. D. L. (2006) The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell* **9**, 333–339.

34. Wadlow, R., and Ramaswamy, S. (2005) DNA microarrays in clinical cancer research. *Curr. Mol. Med.* **5**, 111–120.
35. The Tumor Analysis Best Practices Working Group. (2004) Expression profiling—Best practices for data generation and interpretation in clinical trials. *Nat. Rev. Genet.* **5**, 229–237.
36. Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175.
37. Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., et al. (2003) Source: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31**, 219–223.
38. Kent, W. J. (2002) BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
39. Smit, A., Hubley, R., and Green, P. (1996–2004) RepeatMasker Open 3.0. http://www.repeatmasker.org/.
40. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159.
41. Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47.
42. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34**, D173–D180.
43. Murray, C. G., Larsson, T. P., Hill, T., Bjorklind, R., Fredriksson, R., and Schioth, H. B. (2005) Evaluation of EST-data using the genome assembly. *Biochem. Biophys. Res. Commun.* **331**, 1566–1576.
44. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58.
45. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517.
46. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29.
47. Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., et al. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* **2**, 351–356.
48. Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., and Quackenbush, J. (2005) Independence and reproducibility across microarray platforms. *Nat. Methods* **2**, 337–344.

49. Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., et al. (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350.
50. Simon, R. M., and Dobbin, K. (2003) Experimental design of DNA microarray experiments. *Biotechniques* Suppl, 16–21.
51. Kerr, M. K., and Churchill, G. A. (2001) Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.
52. Dobbin, K., Shih, J. H., and Simon, R. (2003) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes *J. Natl. Cancer Inst*. **95**, 1362–1369.
53. Cox, W. G., and Singer, V. L. (2004) Fluorescent DNA hybridization probe preparation using amine modification and reactive dye coupling. *Biotechniques* **36**, 114–122.
54. Virtanen, C., Ishikawa, Y., Honjoh, D., Kimura, M., Shimane, M., Miyoshi, T., et al. (2002) Integrated classification of lung tumors and cell lines by expression profiling. *Proc. Natl. Acad. Sci. U S A* **99**, 12357–12362.
55. Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., et al. (2003) Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol*. **10**, 119–142.
56. Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* **7**, 106.
57. Tsai, C.-A., Wang, S.-J., Chen, D.-T., and Chen, J. J. (2005) Sample size for gene expression microarray experiments. *Bioinformatics* **21**, 1502–1508.
58. Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. **5**, R80.
59. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet*. **32**(Suppl), 496–501.
60. Smyth, G. K., Yang, Y. H., and Speed, T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol*. **224**, 111–136.
61. Weiner, A. M. (2002) SINEs and LINEs: the art of biting the hand that feeds you. *Curr. Opin. Cell Biol*. **14**, 343–350.
62. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., et al. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet*. **14**, 457–460.
63. Yang, I., Chen, E., Hasseman, J., Liang, W., Frank, B., Wang, S., et al. (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*. **3**, R0062.
64. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A* **95,** 14863–14868.

65. Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001) Validating clustering for gene expression data. *Bioinformatics* **17**, 309–318.
66. Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455–466.
67. Cui, X., and Churchill, G. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4**, 210.
68. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discover rate: a practical and powerful approach to multiple testing. *J. Royal Stats. Soc.* **57**, 289–300.
69. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U S A* **98**, 5116–5121.
70. Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., et al. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **1**, R0003.
71. Rajeevan, M. S., Vernon, S. D., Taysavang, N., and Unger, E. R. (2001) Validation of Array-based gene expression profiles by real-time (Kinetic) RT-PCR. *J. Mol. Diagn.* **3**, 26–31.
72. Beissbarth, T., and Speed, T. P. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465.
73. Zhong, S., Tian, L., Li, C., Storch, K. F., and Wong, W. H. (2004) Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. *Proc. IEEE. Comput. Syst. Bioinform Conf.* 425–435.
74. Bussey, K. J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W. C., Kuo, W. L., et al. (2006) Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Ther.* **5**, 853–867.
75. Mootha, V. K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., et al. (2003) From the cover: identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc. Natl. Acad. Sci. U S A* **100**, 605–610.
76. Segal, E., Wang, H., and Koller, D. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**, i264–i272.
77. van Noort, V., Snel, B., and Huynen, M. A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.* **19**, 238–242.
78. Semon, M., and Duret, L. (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* **23**, 1715–1723.

7

Microarrays—Analysis of Signaling Pathways

Anassuya Ramachandran, Michael A. Black, Andrew N. Shelling, and Donald R. Love

Summary

Microarrays provide a powerful means of analyzing the expression level of multiple transcripts in two sample populations. In this study, we have used microarray technology to identify genes that are differentially regulated in response to activin-treated ovarian cancer cells. We find a number of biologically relevant genes that are involved in regulating activin signaling and genes potentially contributing to activin-mediated growth arrest appear to be differentially regulated. Thus, microarrays are an important tool for dissecting gene expression changes in normal physiological processes and disease.

Key Words: Activin; Affymetrix; bioconductor; microarrays; quantitative real-time RT PCR; signaling pathways.

Abbreviations: cDNA/cRNA – copy or complementary cDNA/RNA; ds DNA – double stranded DNA; GO – gene ontology; qRT-PCR – quantitative real-time RT PCR; SL – signal log ratio

1. Introduction

Gene transcription plays a central role in biology, which is highlighted by the fact that the homeotic genes that are indispensable for initial embryo patterning are transcription factors. Indeed, deregulated transcription is a feature of all cancers, a message that is reinforced by the observation that many transcription factors are altered in cancer. For example, activation of the *C-MYC* proto-oncogenic transcription factor occurs in approximately 70% of all

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

human cancers ***(1)***. Furthermore, the most frequently mutated gene in cancer is *TP53*, which is a well documented transcription factor with tumour suppressive abilities ***(2)***. It is therefore important to invest in technologies that allow the simultaneous and large scale detection of gene transcripts in order to understand normal biological pathways and how they are affected in disease.

The importance of transcriptional control is also highlighted by the number of signaling pathways that ultimately regulate the activity of transcription factors. One such pathway is that mediated by the TGFβ superfamily of secreted ligands. These molecules act *via* cell surface serine/threonine kinase receptors to initiate a signaling cascade from the cell membrane to the nucleus *via* SMAD proteins. The SMADs thus act as both signal transducers and activators of gene expression. It is well documented that the activation of the SMAD cascade by TGFβ leads to a potent anti-mitogenic effect on a variety of cells ***(3)***. We have recently found a similar role for the related molecule activin on ovarian cancer cell proliferation (Ramachandran A. *et.al.*, manuscript in preparation). Therefore, microarray analysis was undertaken to identify global changes in gene expression mediated by activin.

2. Materials

2.1. Cell Culture

1. Alpha Minimal Essential Media (α-MEM) (Invitrogen, Carlsbad, USA).
2. Fetal calf serum (FCS, Invitrogen, Carlsbad, USA).
3. Trypsin (Invitrogen, Carlsbad, USA).
4. Activin A (R&D Systems, Minneapolis, USA) was reconstituted in phosphate-buffered saline supplemented with 0.1% bovine serum albumin to give a final concentration of 10 μg/mL. This solution was stored in 25μL aliquots at –80°C.

2.2. Sample Preparation and RNA Extraction

1. Qiagen QIAShredders (Qiagen, Hilden, Germany)
2. Qiagen RNeasy columns (Qiagen, Hilden, Germany)
3. Beta mercaptoethanol (β-mercaptoethanol) (Riedel-de Haën, Seelze, Germany)

2.3. RNA Quality Assurance

1. Bioanalyzer Lab-on-a-chip assay (Agilent, Palo Alto, USA)
2. Standards for the Lab-on-a-chip assay (Agilent, Palo Alto, USA)

2.4. Microarray Reagents

1. cRNA Labelling Kit (Affymetrix, Santa Clara, USA)
2. GeneAmp® 9700 thermal cycler (Applied Biosystems, Foster City, USA)

3. GeneChip® Hybridization Oven 640 (Affymetrix, Santa Clara, USA)
4. Human HG-Focus Arrays (Affymetrix, Santa Clara, USA)
5. GeneChip® Fluidics Station 450 (Affymetrix, Santa Clara, USA)
6. GeneChip® Scanner 3000 (Affymetrix, Santa Clara, USA)

2.5. cDNA Synthesis and Quantitative RT-PCR

1. Thermoscript Reverse Transcription Kit (Invitrogen, Carlsbad, USA)
2. MJ-Research PTC-100 Thermal Cycler (Bio-Rad, Hercules, USA)
3. Sybr Green PCR MasterMix (Applied Biosystems, Foster City, USA)
4. SDS7700 Sequence Detection System (Applied Biosystems, Foster City, USA)

3. Methods

3.1. Cell Culture for RNA Extraction

1. The epithelial ovarian cancer cell line OVCAR3 was cultured in α-MEM supplemented with 5% FCS in 100-mm tissue culture dishes. Cells were passaged at approximately 80% visual confluence with trypsin.
2. Prior to seeding for activin treatment and RNA extraction, cells were serum starved for 24 h in α-MEM supplemented with 0.2% FCS.
3. For activin treatment and RNA extraction, cells were seeded at a density of 225,000 cells/10-mm dish in α-MEM supplemented with 2.5% FCS (*see* **Note 1**).
4. Recombinant human activin was added to a final concentration of 10 ng/mL.
5. Ten milliliters of culture were used for each dish.
6. Four 10-mm dishes per treatment were pooled at each time point (*see* **Note 2**). At 0 h, one million cells were lysed using Buffer RLT:β-mercaptoethanol (*see* **Note 3**) from the Qiagen RNeasy Kit. The lysate was homogenized by centrifugation through a QIAShredder at maximum speed in a benchtop centrifuge for 2 min.
7. At the other two time points investigated (12 h and 121 h), cells were lysed by adding 650 μL of Buffer RLT:β-mercaptoethanol to one 10-mm culture dish and transferring the lysate sequentially through the remaining three dishes. The lysates were homogenized by centrifugation through a QIAShredder at maximum speed in a benchtop centrifuge for 2 min.
8. After homogenization, total RNA was extracted using the Qiagen RNeasy Kit according to the manufacturer's instructions.
9. Total RNA was eluted in 40μL of RNAse free water, and 1μL of 1U/μL RNAseOUT RNAse Inhibitor (Invitrogen) was added to each eluate.
10. RNA was stored at –80°C.

3.2. Assessment of RNA Quality

1. The integrity of the total RNA was verified using the RNA 6000 Nano Assay from Agilent according to the manufacturer's instructions. The Agilent RNA 6000 Nano Assay uses microfluidic capillary electrophoresis to separate total RNA in a

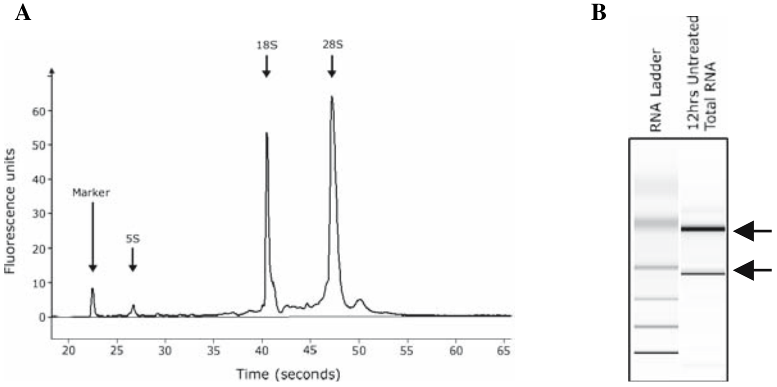


Fig. 1. Agilent RNA 6000 Nano Assay data. **A:** A representative trace of an RNA sample. The trace can be analyzed for a number of features that indicate good quality undegraded RNA, including a uniform and low baseline and the presence of distinct 18S and 28S rRNA peaks at a ratio close to 2. In some instances, the 5S rRNA may also be visible. The Agilent Bioanalyzer software also generates a virtual gel representing predicted band intensities of the 18S and 28S rRNA species (*see* **B**, indicated by arrows), and any RNA degradation, which offers a simple visual representation of the data. A marker for sizing is also shown.

packed gel matrix according to molecular weight. A representative resultant trace is presented in **Fig. 1**.

2. An additional feature of the Agilent 2100 Expert Software is the generation of an RNA Integrity Number (RIN) as an indicator of RNA integrity. The RIN is based on an analysis of all areas of the electropherogram; a RIN of 10 represents a non- degraded RNA sample, while a RIN of 1 represents a sample that is likely to have undergone extensive degradation. All samples used for these microarray experiments had RIN values of at least 9.0.

3.3. RNA Labelling and Array Hybridization

The choice of One Cycle cDNA synthesis or Two Cycle cDNA synthesis is dependent on the amount of starting material. In the case of total RNA concentrations in the range of 1–10 μg, the One Cycle cDNA synthesis can be performed. In the case of lower total RNA concentrations (10–100ng), the Two Cycle cDNA synthesis protocol should be used, as it would lead to sufficient material for further analysis. For the experiments described here, 2.5 μg of total RNA were used as starting material. The One Cycle cDNA synthesis was therefore performed as described in the Eukaryotic Target Preparation guide published by Affymetrix. As such, the entire protocol will not be repeated here, rather attention will be drawn to points of interest in the procedure.

1. The first step in the One Cycle cDNA synthesis involves the preparation of poly-A RNA spike-in controls (*see* **Note 4**). The Poly-A Control Stock must be serially diluted such that the final copy number ratios are 1:100,000, 1:50,000, 1:25,000, and 1:7,500 for *lys*, *phe*, *thr*, and *dap*, respectively.
2. The dilution series for the Poly-A spike in controls are dependent on the amount of starting RNA. For example, 1 μg of total RNA requires a dilution of the stock by 1:20 followed by two serial dilutions of 1:50; 5 μg of total RNA requires a dilution of the stock by 1:20 followed by two serial dilutions of 1:50 and then 1:10. For 2.5 μg of total RNA, the stock is diluted 1:20 followed by two serial dilutions of 1:50 and then 1:35.
3. The second step of the One Cycle cDNA Synthesis involves reverse transcription of the total RNA to single stranded cDNA using a modified oligo(dT) primer (T7-oligo(dT) primer; *see* **Note 5**). All components are assembled on ice and amplified as described by Affymetrix.
4. The third step of the One Cycle cDNA Synthesis involves generating a double-stranded (ds) cDNA molecule with the use of *E.coli* DNA ligase, *E.coli* DNA polymerase I, and RNase H. The ds cDNA is made blunt ended with T4 DNA polymerase. All components are assembled on ice and amplified as described by Affymetrix. The ds cDNA is then cleaned up with the Sample Cleanup Module from Affymetrix according to the manufacturer's instructions.
5. The next step in the protocol involves the generation of biotin-labeled cRNA from the ds cDNA template. This is performed using the in vitro transcription (IVT) labeling kit from Affymetrix, according to the manufacturer's instructions. This reaction is assembled at room temperature as the spermidine in the buffer can lead to precipitation of ds cDNA. The principle behind this step is that the T7 RNA polymerase generates a cRNA molecule that is labelled with a biotinylated pseudouridine molecule.
6. The labeled cRNA is then cleaned with the Sample Cleanup Module from Affymetrix according to the manufacturer's instructions.
7. The labeled cRNA is quantified by diluting 1:100 with RNAse free water and analyzed by spectrophotometric analysis using a Nanodrop.
8. The amount of cRNA produced from the *in vitro* transcription is calculated by subtracting the amount of RNA added to the IVT labeling reaction (based on the amount of input ds DNA) from the final yield of nucleic acid as assayed by the Nanodrop.
9. Fifteen micrograms of labeled cRNA (*see* **Note 6**) are then fragmented using the Fragmentation buffer from Affymetrix according to the manufacturer's instructions.
10. Ten micrograms of the fragmented cRNA are used to prepare the target hybridization cocktail (*see* **Note 7**). This cocktail also contains the control

B2 oligonucleotide for grid alignment, and the eukaryotic hybridization control (*see* **Note 8**). The cRNA is hybridized onto HG-FOCUS arrays as described by Affymetrix in a Hybridization Oven 640, for 16 h (overnight) at 45°C.

11. After overnight hybridization, the arrays are washed using the GeneChip® Fluidics Station 450 and read on a GeneChip® Scanner 3000.

3.4. Microarrays

Most small-scale microarray experiments are concerned with detecting genes that undergo significant changes in expression between two experimental conditions. In terms of experimental design, Affymetrix covers many of the technical aspects in their "Data Analysis Fundamentals" manual, in particular, technical *versus* biological replication, and RNA pooling. Here we focus on a simple comparison between two experimental conditions (treated and untreated), at two post-treatment time points (12 h and 121 h). A single biological replicate was available from each condition-time point combination, giving a total of four arrays. The HG-FOCUS arrays described here were produced by Affymetrix.

3.4.1. Quality control

The first step in analyzing microarrays involves a quality control assessment using the controls that are introduced at various stages during the target preparation.

1. Proper grid alignment of the array is undertaken by using the hybridization signal of the B2 oligonucleotide in order to ensure accurate probe set assignment.
2. If an accurate grid alignment is not achieved, manual alignment of the grid must be undertaken. Upon hybridization, the B2 oligonucleotide should allow the identification of the chip name on the array.
3. Additional eukaryotic hybridization controls are the *bioB*, *bioC*, *bioD*, and *cre* genes that are added in the hybridization cocktail. These controls are present at final concentrations of 1.5pM, 5pM, 25pM, and 100pM, respectively. At 1.5pM, the *bioB* transcript is present at the limit of assay sensitivity. The 3' probe set for the *bioB* transcript was called as present in 2 of 4 arrays, marginal in 1 of 4 arrays and absent in 1 of 4 arrays; these calls were in the accepted range suggested by Affymetrix. All other transcripts were called as present with increasing signal intensities reflecting the increasing molar concentrations of the transcripts.
4. Hybridization of the poly-A spike-in mRNA is also inspected. The 3' end of all the four spike-in controls are detected on the arrays, and in order of increasing signal intensity; *lys*, *phe*, *thr*, and *dap* were called as present.
5. The HG-FOCUS arrays also have internal controls genes (β-actin and GAPDH) that allow the assessment of data quality. The ratio of the 3' probe set signal to the 5' probe set signal for these genes should not be more than three, for a one cycle cDNA synthesis.

3.4.2. Microarray data analysis

We undertook the analysis of microarray data using two different statistical software packages. The first was the open-source software, Bioconductor ***(4)***, and the second used the MAS5 suite within GCOS (Affymetrix). The MAS5 software is convenient (in that it is bundled with the software for the Affymetrix system) and a relatively user-friendly option for data analysis, but the methodology is somewhat outdated, and does not take advantage of many of the advances in microarray data analysis that have been made over the past few years. In this respect, use of the Bioconductor software provides researchers with access to data analysis tools that are constantly being improved and updated by an international group of developers.

3.4.2.1. Microarray Data Analysis Using R

Bioconductor is an open source software project designed to develop tools for the analysis of genomic data ***(4)***. Bioconductor extends the functionally of the R computing environment ***(5)*** by providing packages that contain functions for the analysis of specific types of data. Although these packages are not limited to the analysis of data from microarray experiments, the current popularity of this technology has meant that a large number of the available Bioconductor packages are devoted to microarrays. The Bioconductor packages available for the analysis of Affymetrix microarray data provide users with extremely powerful statistical methods, many of which have become the *de facto* standard for Affymetrix data analysis.

1. In order to utilize the functionality offered by the Bioconductor project, it is necessary to download and install the R software package. This is available for download from the main CRAN (Comprehensive R Archive Network) website (www.cran.r-project.org). Pre-built versions of the R software are available for all major computing platforms, including Windows, Mac OS X, and specific Linux distributions, with source code available for building on Unix-based systems. Once installed, R provides access to detailed html-based help documentation *via* a web browser.
2. Once R has been successfully set up, the necessary Bioconductor packages can be installed by running a simple script which is available from the Bioconductor website (www.bioconductor.org). This allows users to choose either a full (very large) installation of Bioconductor (*see* **Note 9**), or one tailored to their specific needs. An option provides the ability to download a subset of packages which are suitable for the analysis of Affymetrix data.
3. Data from the CEL files are read into R using the Affy package (*see* **Note 10**).
4. Quality control reports are produced using the AffyQCreport and AffyPLM packages (*see* **Note 11**).

5. Probe set summaries are produced using the Robust Multi-chip Analysis (RMA) method *(6)* (*see* **Note 12**).
6. In order to detect probe sets undergoing significant changes in expression between the two experimental conditions at each time point, the EBarrays package *(7)* is used to produce estimates of the probability of differential expression for each gene (*see* **Note 13**).
7. Genes with probabilities of differential expression at either time point of greater than 0.5 are considered to have undergone significant changes (**Fig. 2**).
8. The lists of significant genes from each time point are then used as inputs into the hypergeometric testing functions geneGoHyperGeoTest and geneKeggHyperGeoTest in the Category package in order to investigate over-representation of specific Gene Ontology categories and KEGG pathways, respectively.
9. Within each of these tests, the resultant *p*-values are adjusted for multiple hypothesis testing using the false discovery rate controlling method *(8)*.

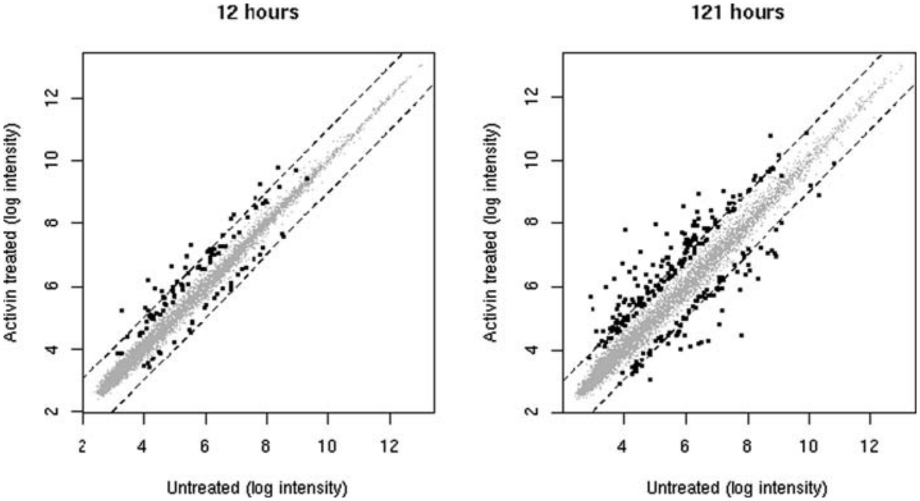


Fig. 2. Activin-treated versus -untreated RMA transformed log intensities for 12-h and 121-h samples. The EBarrays package showed a total of 309 probe sets that were found to be differentially expressed at either time point, with 54 only detected at 12 h, 200 only detected at 121 h, and 55 detected at both time points; significantly differentially expressed genes are marked in black. At both time points, more probe sets are up-regulated in response to activin treatment than are down-regulated (80 up and 29 down at 12 h, 171 up, and 84 down at 121 h). Of the 55 significant probe sets in common between the two time points, 51 were up-regulated at both time points, and 4 were down-regulated. The dotted lines indicate the boundaries for differentially expressed probe sets.

10. In our study, the analysis of Gene Ontology annotation *via* the geneGoHyperGeoTest function reported significant (after adjustment for multiple hypotheses testing) over-representation of the "Development" GO term among the probe sets found to be differentially expressed at 12 h (*P* = 0.015). This was also seen at 121 h (*P* = 0.002), with the "DNA replication" GO term also found to be over-represented in the differentially expressed probe sets (*P* = 0.08) (*see* **Note 14**).
11. The Analysis of KEGG annotation information *via* the geneKeggHyperGeoTest command indicated that probe sets from both the TGF-β and Hedgehog signaling pathways were significantly over-represented at 12 h (*P* = 0.012 and *P* = 0.052, respectively) (*see* **Note 15**).
12. Genes from the TGF-β signaling pathway were again found to be over-represented at 121 h (*P* = 0.039), with 6 of the 11 differentially genes expressed up-regulated, and five down-regulated (total pathway size, 73 genes).
13. Genes from the intracellular matrix (ECM) receptor interaction pathway were also found to be over-represented at 121 h (*P* = 0.0439), with all 10 differentially expressed genes up-regulated (total pathway size, 60 genes).

3.4.2.2. Microarray Data Analysis Using MAS5

Detailed information about Affymetrix software (GCOS) is provided on their website (www.affymetrix.com/support/technical/manual/expression_manual.affx), along with a supplementary manual relating to data analysis (www.affymetrix.com/Auth/support/downloads/manuals/data_analysis_fundamentals_manual.pdf) (*see* **Note 16**).

1. The background of all slides assayed here was under 55 (arbitrary fluorescent units).
2. Prior to the comparison of pairs of arrays, global scaling to an average target intensity of 150 can be applied to all slides, resulting in scaling factors ranging from 1.78 to 2.23 (*see* **Note 17**).
3. For comparative analysis, untreated samples were used as baseline chips against which the activin-treated samples at each time point were compared. All tunable parameters were left at default values.
4. Two algorithms are used to compute the significance of the comparative changes in gene expression.
5. The first algorithm, termed the change algorithm, generates a change *P*-value and an associated change call (increased, marginally increased, no change, marginally decreased or decreased). The change *P*-value is calculated using the Wilcoxon signed rank test.
6. The second algorithm, termed the signal log ratio (SLR) algorithm, calculates the magnitude of the change, and independently generates the direction of change (that is, positive or negative corresponding to increased or decreased, respectively). The magnitude of the change is calculated using the one-step Tukey's Biweight method, and is presented on a log scale to the base 2 (**Fig. 3A**). Thus, an SLR of

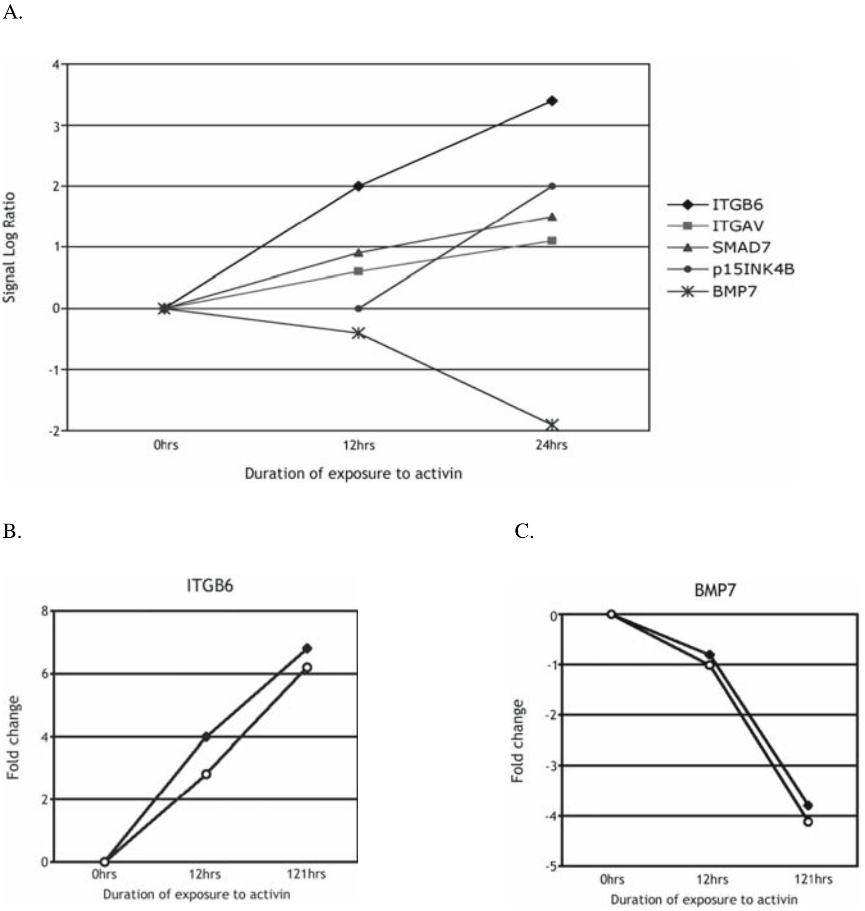


Fig. 3. MAS 5 generated signal log ratios (SLRs) of selected transcripts from activin-treated cells. SLRs of some biologically relevant targets that were identified as differentially regulated by microarray analysis of OVCAR3 cells after treatment with activin are shown in **A**. The integrins α5 and β6 (ITGAV and ITGB6, respectively) are extracellular matrix interacting genes that are involved in the activation of latent TGFβ to its active form. They may be important in establishing a positive feed-forward loop of TGFβ/activin signaling in these cells. BMP which competes activin for some intracellular SMAD targets is down-regulated in response to activin in OVCAR3, potentially allowing enhanced activin signaling to occur. SMAD7 is known to be induced in response to activin treatment in other cell types and is likely to regulate the extent of activin signaling in these cells. The induction of the cell cycle inhibitor p15INK4B is likely to contribute to activin-mediated growth arrest in OVCAR3 cells. The fold changes in expression levels of ITGB6 **(B)** and BMP7 **(C)** were assayed by

1 represents a two fold increase in transcript and an SLR of –1 represents a two fold decline in transcript abundance.

7. Probe sets that gave absent calls for both activin-treated and untreated samples were not analyzed. If untreated samples had an absent call, but treated samples had a present call, these data were included as they demonstrated an increase in gene expression in response to activin treatment (and vice versa).
8. Data are first sorted based on the change algorithm call; only data that are called as increased or decreased are analyzed further. Of these, only probe sets with a SLR of ≥ 1 (twofold induction) or ≤ -1 (twofold suppression) were included in this study.
9. Genes identified as differentially regulated can be uploaded into the NetAffx analysis centre from Affymetrix, and genes are analyzed based on GO-Browser annotation and literature searches.
10. In our study, 192 genes were identified as differentially regulated in response to activin; 88 of these genes were detected at both time points, 12 genes were detected at 12 h and 92 genes were detected at 121 h. Of the genes that were differentially regulated, the majority were up-regulated in response to activin treatment (152 up-regulated genes versus 40 down-regulated genes).

3.5. Validation of Transcript Level Changes

To ensure accurate validation of the microarray data, RNA samples that are hybridized onto the arrays are used for the subsequent quantitative real-time RT PCR (qRT-PCR) validation of target genes.

3.5.1. Isolation of RNA and Reverse Transcription

1. Six micrograms of RNA are pre-treated with 1.5 units of DNAse I (Invitrogen) according to manufacturer's instructions.
2. The DNAse I-treated RNA is then divided equally into three aliquots: two of which are reverse transcribed with a reverse transcriptase (RT+) and one of which is not

Fig. 3. (Continued) qRT-PCR (open circles) and MAS 5 (closed squares). Fold changes at 12 h and 121 h represent the expression level of the transcripts in response to activin treatment relative to the expression in the untreated sample at that time point. Fold changes obtained by qRT-PCR correlate well with fold changes of these transcripts identified by MAS 5. Note that the 0-h sample represents RNA from cells at the beginning of the experiment; therefore SLRs and fold changes do not apply as there is only one sample at this time point. It is depicted in these graphs to demonstrate the magnitude and direction of changes in gene expression that occur over time in response to activin treatment.

reverse transcribed (RT−) (*see* **Note 18**). To each aliquot, a mixture containing oligo$(dT)_{20}$ primers (2.5 μM), dNTPs (1 mM of each dNTP) and RNAse free water to a final volume of 12 μL is added.

3. The mixture is heated to 65°C for 10 min and then immediately placed on ice.
4. A cDNA synthesis master mix is then assembled and added to the RNA/primer/dNTP mix to a final volume of 19 μL. The cDNA synthesis master mix contains the cDNA Synthesis Buffer (final concentration of 1×), DTT (final concentration of 5 mM) and RNAseOUT RNAse inhibitor (2 units).
5. This mixture is incubated at 55°C for 5 min and then placed on ice.
6. To the RT+ reactions, 1 μL Thermoscript RT (15U/μL) (*see* **Note 19**) is added, while to the RT- reaction, 1 μL DEPC water is added.
7. This reaction is incubated on a preheated block at 55°C for 1.5 h. The reaction is then terminated by incubating at 85°C for 5 min.
8. Finally the RNA template in the RNA:cDNA hybrid is degraded by incubating at 37°C for 20 min.
9. This cDNA is diluted by adding 80 μL of sterile, milliQ water.

3.5.2. Quantitative Real-Time RT PCR (qRT-PCR)

The expression of biologically relevant genes chosen from the arrays was analyzed in OVCAR3 at the transcript level by quantitative real-time reverse transcription PCR (qRT-PCR) (**Fig. 3B,C**). The data analysis for normalization and relative quantification of gene expression across the samples (that is, OVCAR3 +/− activin at various time points) was performed essentially as described earlier ***(9)***, and is outlined below.

1. PCR primers are designed using Primer Express software (Applied Biosystems), with all products being less than 150 base pairs (bp) (*see* **Note 20**).
2. All the samples for a given primer pair are analyzed on the same plate, with each sample analyzed in triplicate.
3. Each qRT-PCR is performed in a final volume of 10 μL, with 1× Sybr Green Master Mix (Applied Biosystems), 0.1 U uracil DNA glycosylase (Invitrogen), 300 nM of each primer and 2 μL of diluted cDNA (above).
4. The amplification efficiency, designated E, of each primer pair is calculated in the linear phase of amplification of each qRT-PCR using the LinRegPCR applet for Excel ***(10)***, and averaged across the plate.
5. The average amplification efficiency is used to calculate the "raw" quantity of each transcript across the samples based on their C_T values, which corresponds to the fractional cycle number at which a given amplification reaction crosses a defined threshold ($E^{(\mathrm{Min}C_T - C_{T,X})}$, where $C_{T,X}$ represents the C_T value of each qRT-PCR; Applied Biosystems SDS Chemistry Guide).

6. The two most stable housekeeping genes, beta actin (*ACTB*) and the TATA box binding protein gene (*TBP*), were used to calculate the normalization factor (NF) for each sample using the geNorm Excel VBA applet ***(9)***.
7. The mean raw quantities for each sample are divided by the NFs to generate the relative abundance of each transcript in that sample.
8. Standard deviations are calculated using the error propagation rules for independent variables as described in the geNorm User Manual ***(9)***
9. The abundance of each transcript in the untreated sample at each time point was used as the calibrator (that is, expression level = 1), and expression levels in the activin-treated samples are represented by reference to this value.
10. Statistical analyses are performed using the student t-test, 2-tailed.

4. Notes

1. Confluence of cells can alter their responsiveness to secreted cytokines. For OVCAR3, plating 225,000 cells/10-mm dish in 2.5% FCS ensures that they are still actively proliferating after 5 days in culture. However, the appropriate plating density has to be determined for each individual cell line.
2. The decision to pool samples is one that requires careful consideration before undertaking a microarray experiment. In many studies, pooling of samples will lead to a loss of data due to target dilution and variability in gene expression from different samples. However, *in vitro* cell culture tends to generate a relatively homogenous cell population compared to *in vivo* tissues; thus, the decision was made to pool four 10-mm culture dishes for each treatment point for OVCAR3 to increase the concentration of starting total RNA.
3. β-Mercaptoethanol is a potent reducing agent that disrupts disulphide bonds in RNAses, thereby inhibiting their activity. However, β-mercaptoethanol can be oxidised and once this occurs RNAses regain their activity. Thus, β-mercaptoethanol must be added to fresh buffer RLT and this should be used within 24 h.
4. The Poly-A spike in RNA are *in vitro* synthesized *lys*, *phe*, *thr*, and *dap* polyadenylated gene transcripts from the bacterium *Bacillus subtilis* that are absent in all eukaryotic samples. These serve as a quality control to determine the efficiency of the *in vitro* transcription and labelling procedures as they are added to the sample RNA and handled along with it.
5. The modified T7-oligo(dT) primer possesses a T7 RNA polymerase binding site at the 5’ end of the primer that is subsequently used in the generation of cRNA.
6. The amount of cRNA to be fragmented depends on the format of the Affymetrix array format. The HG-FOCUS arrays are 100 format arrays (midi-format arrays); thus, 15 μg of cRNA are fragmented. For 49/64 format arrays (standard arrays), 20 μg of cRNA are fragmented.
7. The amount of cRNA hybridized onto an array also depends on the array format. For the HG-FOCUS arrays (midi-format arrays), 10 μg of fragmented cRNA are

hybridized. For the 49/64 format arrays (standard arrays), 15 μg of fragmented cRNA should be hybridized.

8. The eukaryotic hybridization controls added after the *in vitro* transcription and fragmentation of the cRNA controls for hybridization efficiency of the samples onto the arrays.
9. The Bioconductor packages contain documentation (including example commands for the analysis of data) in the style of a tutorial, which can be viewed as a PDF, or used to load example commands directly into R. There is a need to have a chip-specific annotation package installed before trying to normalize Affymetrix data using the Affy package from Bioconductor.
10. The CEL file contains intensity information for each probe on the array. Sequences are represented by multiple perfect-match (PM) and mis-match (MM) probes (probe sets), and genes can be represented by more than one probe set.
11. These reports are used to identify problematic arrays, and the investigator can use this information to decide whether questionable arrays should be included in the analysis.
12. When performing statistical analysis, it is important to take the number of probe sets representing each gene into account. In the over-representation analysis this can be most easily accomplished by merging the data across all probe sets which represent the same gene; for example, by taking the median value across the probe sets for that gene.
13. To use most statistical methods (and obtain meaningful results), biological replication is required, as replicates are necessary for the estimation of per-gene variability. If replicate data are not available, then methods which assume a constant coefficient of variation, i.e., constant variance on the log scale, such as those available in the EBarrays package, can be used to obtain information about differential expression. If technical replicates are to be used in conjunction with biological replications, then the correlation between the technically replicated arrays e.g. a pair of arrays to which the same sample was hybridized, needs to be accounted for in the statistical analysis. Currently the limma package ***(11)*** in Bioconductor is one of the few methods able to do this correctly.
14. Many of the methods for investigating over-representation of function categories can also be accessed *via* web interfaces. In some cases this approach may be preferable to using the Bioconductor annotation packages, as results are generally formatted as HTML tables, and annotation information is likely to be updated more regularly. Examples of such resources include DAVID at NCBI, and the NetAffx service provided by Affymetrix.
15. While five of the 73 genes in the TGF-β signaling pathway were differentially expressed (three up, two down), only two (one up, one down) of the 38 genes in the Hedgehog signaling pathway were differentially expressed. It should be noted that while entire pathways may not be significantly altered, individual components

of the pathway may be biologically relevant and, thus, differentially regulated as seen here.

16. Although the Affymetrix GCOS software provides a number of methods for the analysis of data from both single and multiple array experiments, these methods are relatively simple, and do not reflect the extensive range available for the analysis of data from microarray experiments.
17. Prior to the comparison of two arrays with MAS5, scaling or normalization must be applied to the arrays. With global scaling, the signal intensity of each array is scaled to reach a user-defined threshold either using all probe sets or a defined probe set determined in the mask file. In the case of normalization, the signal intensity of one array (the experimental array) is normalized to the signal intensity on the other array (the baseline array). For experiments where the majority of the transcripts are expected to be unchanged, Affymetrix recommend the use of global scaling, and as such this was carried out for these experiments. Scaling factors should be within three orders of magnitude of each other; greater than this suggests technical errors in the analysis.
18. RT+ reactions represent the cDNA synthesis reactions while the RT- reactions control for genomic DNA contamination in the RNA samples.
19. If less than 1 ng of RNA is used as starting material, 7.5U of Thermoscript should be used for each reaction.
20. Primers for qRT-PCR were designed against the target sequences covered by the Affymetrix probe sets on the HG-FOCUS arrays.

Acknowledgments

The authors acknowledge the financial support of the University of Auckland Research Committee, the Lottery Grants Board of New Zealand, the Maurice and Phyllis Paykel Trust and the Cancer Society of New Zealand.

References

1. Nilsson, J. A., and Cleveland, J. L. (2003) Myc pathways provoking cell suicide and cancer. *Oncogene* **22**, 9007–9021.
2. Szymanska, K., and Hainaut, P. (2003) TP53 and mutations in human cancer. *Acta Biochim. Pol.* **50**, 231–238.
3. Massague, J. (1998) TGF-beat signal transduction. *Annu. Rev. Biochem.* **67**, 753–791.
4. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.
5. Ihaka, R., and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stats* **5**, 299–314.

6. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.
7. Newton, M. A., and Kendziorski, C. M. (2003) Parametric empirical bayes methods for microarrays, in *The analysis of Gene Expression Data: Methods and Software* (Parmigiani, G., Garrett, E. S., Irizarry, R., and Zeger, S. L., eds.), Springer Verlag, New York.
8. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300.
9. Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, R34.
10. Ramakers, C., Ruijter, J. M., Deprez, R. H., and Moorman, A. F. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66.
11. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 3.

8

Microarrays—Identifying Molecular Portraits for Prostate Tumors with Different Gleason Patterns

Alexandre Mendes, Rodney J. Scott, and Pablo Moscato

Summary

We present in this chapter the combined use of several recently introduced methodologies for the analysis of microarray datasets. These computational techniques are varied in type and very powerful when combined. We have selected a prostate cancer dataset which is available in the public domain to allow for further comparisons with existing methods. The task is to identify biomarkers that correlate with the clinical phenotype of interest, i.e., Gleason patterns 3, 4, and 5. A supervised method, based on the mathematical formalism of *(α,β)–k–feature sets* ***(1)***, is used to select differentially expressed genes. After these "molecular signatures" are identified, we applied an unsupervised method *(a memetic algorithm*) to order the samples ***(2)***. The objective is to maximize a global measure of correlation in the two-dimensional display of gene expression profiles. With the resulting ordering and taxonomy we are able to identify samples that have been assigned a certain Gleason pattern, and have gene expression patterns different from most of the other samples in the group. We reiterate the approach to obtain molecular signatures that produce coherent patterns of gene expression in each of the three Gleason pattern groups, and we analyze the statistically significant patterns of gene expression that seem to be implicated in these different stages of disease.

Key Words: microarray data analysis, prostate cancer, pathway analysis, Gleason score, memetic algorithm, *(α,β)–k–feature set*.

1. Introduction

For the diagnosis and treatment of prostate cancer the Gleason scoring system ***(3)*** is used as a standard approach for clinical decision-making. It takes into account the heterogeneous nature of prostate cancer by combining in a score the

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

two most prevalent features of the tumour. The most common pattern/feature is first given an individual score on a scale from 1 to 5, and to that is added the score of the second most common feature, resulting in a total score that ranges between 2 and 10. Several studies support the thesis that there is a clear relationship between the score and clinical outcomes (*see* **Note 1**).

Recently, gene expression changes have been associated with Gleason scores ***(4)*** and suggest that there are specific changes associated with the different stages of prostate cancer. There are, however, clear difficulties encountered in discriminating between a Gleason score of 4 and 5, signaling that the data analysis was not precise enough to differentiate higher grades of disease. To define prostate cancer more effectively gene expression changes need to be assigned to the apposite Gleason score. However, using some bioinformatic approaches this has proved very difficult to achieve due to the heterogeneity of prostate tumors.

True and colleagues ***(4)*** have noticed how difficult it is to find good gene sets that can act as biomarkers to identify samples as being either Gleason pattern 4 or 5. In their own words: "*We were unable to identify a cohort of genes that could distinguish between patterns 4 and 5 cancers with sufficient accuracy to be useful, suggesting a high degree of similarity between these cancer histologies or substantial molecular heterogeneity in one or both of these groups"*. As our approach to the identification of molecular signatures will try to maximize the similarity of the groups, our methods may help to classify more adequately subsets of this disease into their respective Gleason scores, and thereby provide an accurate gene feature set that can be used to distinguish between the different Gleason grades.

The categorization of molecular events that underlie prostate cancer development and progression has been difficult. This is a direct result of the heterogeneous nature of the tumors themselves. Nevertheless, prior to gene expression analysis there were a number of consistent observations made about some of the molecular changes associated with disease development. These included the large number of somatic mutations identified in the tumors, as well as recognition that growth factors and the hormonal milieu significantly contribute not only to disease development but also its progression. Two factors are of particular relevance to disease risk and these are testosterone and insulin-like-growth factor 1 (IGF-1).

Alterations in the androgen receptor have been linked to a worse prognosis of disease [for review of the role of androgens and the androgen receptor, *see* ***(5)***]. Androgen action results in a series of gene expression changes that affect cellular proliferation by directly increasing expression of genes such as *TMPRSS2:ETV1* fusion gene, *TMPRSS2*, *ERG*, *WISP-2* and indirectly via *Src/Raf-1/MEK1* and *IGF-1*. Androgens are also intimately involved in cell survival as they have been shown to influence the expression of caspase-2, *c-FLIP*, *P53*, *MDM2*, *Hox5a* and *Egr-1*.

The *IGF signaling pathway* appears to be significant as it orchestrates a variety of responses in concert with the androgen receptor. Higher circulating levels of *IGF-1* have been associated with an increased risk of prostate cancer ***(6,7)*** via, at least in part, trans-activation of the androgen receptor pathway through the *IGF-1* receptor, which results in the potentiation of AR signaling (for review *see* **8**).

Studies aimed at identifying other molecular changes associated with prostate cancer have revealed a number of genes that are either lost or have altered gene expression compared to normal tissue. These include genes such as *NKX3-1*, a homeobox gene with prostate specific expression ***(9)***; *PTEN*, a lipid phosphatase that dephosphorylates phosphatidylinositol-3,4,5-tri phosphate, which results in perturbation of the AKT pathway ***(10)***; *CDKN1B* (p27) a cell cycle inhibitor that interacts with the AKT pathway and potentiates loss of *PTEN* ***(11)***; *ATFB1*, a protein that cooperates with *MYB* ***(12)***; *KLF6*, a Kruppel-like zinc finger transcription factor ***(13)***; *ERG* and *ETV*, ETS transcription family members that form fusion proteins with *TMPR22S* ***(14)***.

2. Mathematical Methodology and Approaches

We have described elsewhere the mathematical models and algorithms employed for the identification of the molecular signatures to be discussed below. We consider this work the last in a trilogy of chapters we have published with *Humana Press*. The previous two chapters in this series ***(15,16)*** explain in full detail the mathematical models and their application in finding molecular signatures in Alzheimer's and Parkinson's disease studies. Our approaches rely on the application of combinatorial optimization models and algorithms for their solution (***1,2,17–19***).

Our objective in this chapter is to be complementary to that presented previously and give emphasis to the discussion of the application of these methods for a case study in clinical bioinformatics. We refer the reader to these references for an introduction to these methods. More details of other publications related to our work can be found at the Newcastle Bioinformatics Initiative's website (http://www.cs.newcastle.edu.au/~nbi).

The selection of the problem dataset has allowed us to point to some of the current best practices in microarray experimental design, pre-processing and analysis with statistical methods. We refer to ***(4)*** for the discussion of these important steps and relevant literature. This "shortcut" allowed us to concentrate on the outcomes of our methodologies and their application in the clinical bioinformatics arena.

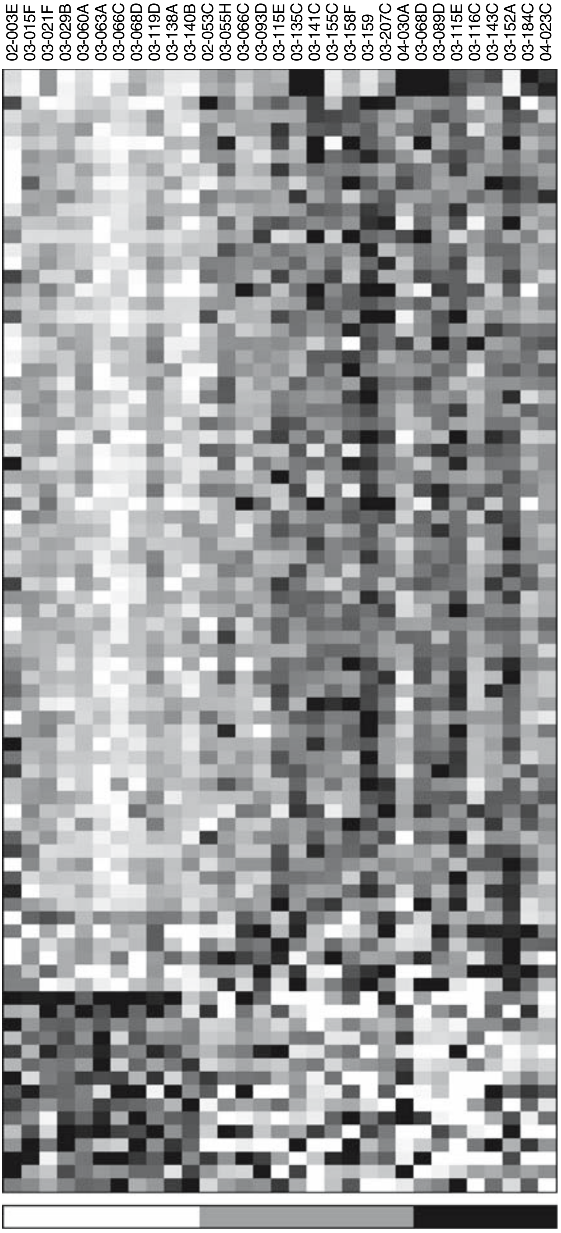
Initially, we verified the results reported in *(4)* and replicated the molecular signature for Gleason pattern 3. This was achieved by selecting the most representative expression patterns of the probes in their gene set (**Fig. 1**). For creating the molecular signatures we have undertaken some very simple preprocessing. We have used 31 samples from the dataset in *(4)* that had been assayed with the GPL3834 (FHCRC Human Prostate PEDB cDNA Array v4) platform (with 15,488 probes) and which originally consisted of 32 samples. We disregarded sample 02-209C from our analysis as the data were acquired using a different gene expression analysis platform and contained fewer probes. After removal of probes that have six or more missing values, there were 14,499 probes and 31 samples to analyze.

2.1. Analysis of Differentially Expressed Genes

True and colleagues *(4)* highlighted the problems their classifier had on an independent set. In their own words: "*Of the 12 cancers histologically called Gleason pattern 3, all but one was correctly classified. Of the cancers with a histological classification of 4 or 4+5, 6 of 11 were correctly classified. As expected, microdissected samples recognized to contain mixed grades of 3+4 were divided between pattern 3 or pattern 4 molecular categories. These results suggest that pattern 3 cancer exhibit relatively consistent molecular alterations, whereas cancers with histological features of patterns 4 and 5 are more diverse and, in some cases exhibit molecular features common to pattern 3 cancers.*"

It seems that, in part, the problems are related to a direct consequence of the approach described in *(4)* since tumors of patterns 4 and 5 had to be grouped to produce their statistically based molecular signature for Gleason pattern 3. To compare our combinatorial optimization approach, we have created three molecular signatures that we denote as Gleason-3-versus-(4+5),

Fig. 1. True and colleagues *(4)* have recently reported that the expression of 86 genes help to distinguish Gleason pattern 3 (in white) from Gleason pattern 4 (gray) and 5 (black). Here we depict the values corresponding to those genes, which are in closer agreement with their reported signature. Grayscale intensities have been selected to make a clear picture of individual gene variation and we have used the same settings in the other figures. Columns correspond to samples, presented here in the same order as they appeared in their figure *(4)*.



GLEASON 3	GLEASON 4	GLEASON 5
LOW GRADE	HIGH GRADE	

Gleason-4-versus-(3+5), and Gleason-5-versus-(3+4). As expected, the samples of Gleason 3 are more coherent than the other two groups. In Gleason-3-versus-(4+5), each pair of samples belonging to different target groups, e.g., one in Gleason 3 and the other that is not, have at least 267 genes differentially expressed, while the within-class similarity is also high (for any pair of samples belonging to the same class we have at least 197 genes with a similar expression). The total number of genes in the signature is 522. We refer to such a signature as an (267,197)-522-feature set. Analogously, molecular signatures for Gleason-4-versus-(3+5) and Gleason-5-versus-(3+4) correspond to (82,92)-187 and (122,99)-223 feature sets (**Fig. 2**).

2.2. Biomarkers of Interest—Gleason Pattern 3

Our genetic signature for Gleason-3-versus-(4+5) has revealed a number of differentially expressed genes in concordance with ***(4)***. In particular, *MAOA (Monoamine oxidase A)* and *DAD1 (Defender against cell death 1)* are in general down-regulated in tumors corresponding to Gleason pattern 3. Immunohistochemical analysis was also performed by True and colleagues ***(4)***, which confirmed their array analysis; *MAOA* protein levels were assessed by immunohistochemical analysis on panels of tissue microarrays (889 cancerous and 469 benign samples). They found that protein expression was elevated in cancerous epithelium relative to benign secretory epithelium and that *MAOA* expression is significantly higher in Gleason 4 and 5 samples in comparison with Gleason 3. In both cases the P-value was lower than 0.0001 (proportional odds-regression analysis). Our analysis has also revealed that to define more accurately Gleason 3 we need to include three other genes. Two of them should be highly expressed *KLF6 (Kruppel-like factor 6)* and *MYBPC1 (Myosin binding protein C, slow type)*, while the other gene, *SPON2 (Spondin 2, extracellular matrix protein)* ***(20,21)*** is down-regulated. We will return to these markers after we discuss the biomarkers for the other two patterns.

2.3. Biomarkers of Interest—Gleason Pattern 4

As expected, it is a real challenge to find markers for this group for the reasons discussed above. We concentrated our attention on four genes that seem to be notoriously down-regulated in most, but not all the samples in the group labelled Gleason 4. The genes are *CRABP2* (*cellular retinoic acid binding protein*), *TPM2* (*tropomyosin 2 beta*), *EDNRA* (*endothelin receptor type A*), and *CTGF* (*connective tissue growth factor*).

CRABP2 is a regulator of anti-carcinogenic activities of retinoic acid and it has been suggested previously that *CRABP2* is down-regulated in prostate

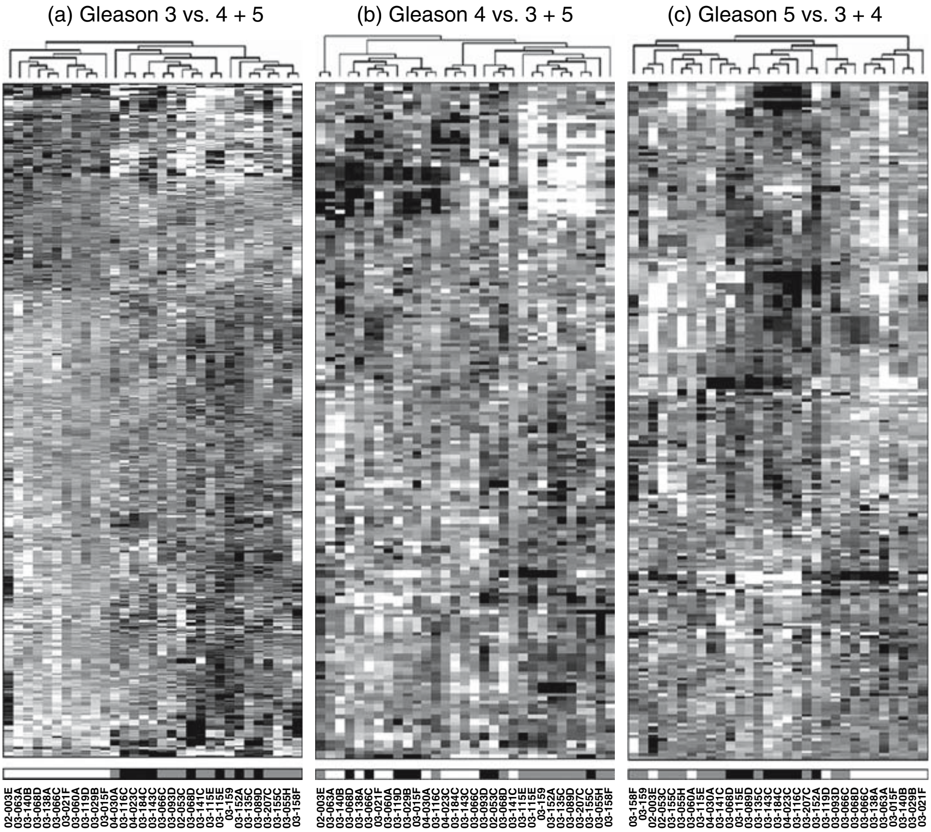


Fig. 2. Molecular signatures that distinguish, in each case, one of the Gleason patterns from the other two. Rows represent genes and columns represent samples (white for Gleason 3, gray for Gleason 4, and black for Gleason 5, as in Fig. 1). We have used the memetic algorithm of Moscato et al. (2006) to find optimal orders of samples and genes. **a:** Samples with Gleason pattern 3 have more than 60 genes up regulated in comparison with the samples of other patterns (**upper left corner**). Sample 02-003E, the only one that expresses *MAOA* and *Death-associated protein (DAP,* BM910328), seems an outlier in the Gleason 3 pattern group. **b**: A molecular signature to identify samples of Gleason pattern 4 (in gray); samples 03-060A, 02-003E, and 03-115E appear to belong to Gleason 4. **c:** Sample 03-135C in this molecular signature for Gleason pattern 5 indicates that its profile is very similar to others of Gleason 5.

cancer ***(22)***. In a study on MCF-7 mammary carcinoma cells ***(23)*** it was observed that retinoic acid treatment of MCF-7 triggered pronounced apoptosis and that *CRABP2* has pro-apoptotic activities (over-expression of *CRABP2* up-regulated *APAF1* and triggered Caspase 7 and Caspase 9 cleavage). *CRAB2* undergoes nuclear localization upon binding of retinoic acid, interacts with the retinoic acid receptor in a ligand dependent fashion, raising the possibility that these prostate tumors (similar to MCF-7) may be retinoic acid resistant ***(24)***.

In this dataset, *CRABP2* is co-regulated with *CTGF*. In stromal tissue, expression of *CTGF* has been linked to the promotion of angiogenesis and prostate cancer tumorigenesis and is a powerful mediator of *TGF-β1* action. In tumor-reactive stroma, *CTGF* expression induced an increase in microvessel density and xenograft tumor growth, suggesting that *CTGF* is a downstream mediator of *TGF-β1* action in cancer-associated reactive stroma, and is likely to be one of the key regulators of angiogenesis in the tumor-reactive stromal microenvironment ***(25)***. Blockage of *CTGF* has been suggested as a therapeutic target against benign prostatic hyperplasia ***(26)***.

Endothelins and their receptors are related to angiogenesis, tumor growth and proliferation, bone metastasis and apoptosis ***(27)***. Gleason 4 samples that over-express *CRABP2* and *CTGF* also have *EDNRA* up-regulated. Over-expression of Endothelin A (*ET-A*) receptor is known to increase with tumor progression. Clinical trials with selective *ET-A* receptor antagonists, such as Atrasentan (ABT-627) are showing promising results ***(28)***. Another study indicated that a combination of *ET-A* antagonists and apoptosis-inducing therapies could be beneficial for prostate cancer ***(29)*** (see also **30**).

TPM2 is also down-regulated in a subset of Gleason pattern 4 tumors. It has recently been shown that *TPM2* may have a role in the loss of actin stress fibers, which in turn is associated with cell transformation and metastasis. Epigenetic suppression of *TPM1* may affect *TGF-β* inducing stress fibers and inhibit cell migration in metastatic cells ***(31)***. As *TGF-β* induction of stress fibers in epithelial cells requires the tropomyosins *TPM1* and *TPM2* genes, it remains to be understood what the effects of *TPM2* down-regulation are in prostate cancer.

2.4. Biomarkers of Interest—Gleason Pattern 5

Analysis of the results produced by our signature for Gleason-5-versus-(3+4) has resulted in an identification of two types of samples within this group. We will centre the discussion on three genes: *CXCR4 (chemokine (C-X-C motif) receptor 4/CD 184 antigen/Fusin), DPP4 (Dipeptidyl-peptidase*

4 / CD26, adenosine deaminase complexing protein 2), and *SPP1 (Secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1*).

The marked over-expression of *CXCR4* is only present in four samples of Gleason type 5. There are also samples with high (but more moderate) expression in three or four samples labeled Gleason 4 and one in the Gleason 3 groups. These samples also over express *ITGB2 (Integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)). CXCR4* is involved in angiogenesis ***(32)*** during tumor invasion and metastasis (***33–38***) and a *CXCL12*-triggered chemoattractive mechanism implicated in tumor cell binding has been uncovered, which has established a connection between chemokine receptor expression and integrin-triggered tumor dissemination ***(39,40)***. *CD164*, *CXCR4*, and *CXCL12* participate together in the localization of prostate cancer cells to bone marrow (***41–43***), and new targeted therapies are being developed to block this process (***44–46***). Kukreja and colleagues ***(47)*** have shown that *CXCL12* induced expression of *CXCR4* in PC-3 cells is dependent on the MEK/ERK signaling cascade and NF-kappa B activation. It has been recently suggested that the shift of *CXCR4* (and *CXCR3*) from the cell surface to the cytoplasm might indicate progression from a low to a highly aggressive phenotype ***(48)***.

DPP4 is specific for luminal secretory cell types of the prostate ***(49)***, and it has been reported to be expressed in prostate cancers and adjacent benign prostatic hyperplastic tissue ***(50)***. *DPP4* is a serine protease with tumor suppressor function, regulating the activities of mitogenic peptides implicated in cancer development. Wesley and colleagues ***(51)*** have shown, *via* silencing *DPP4* with siRNA, an increase in *bFGF* levels and restoration of mitogen-activated protein kinase (*MAPK*)-extracellular signal-regulated kinase (*ERK*)1/2, suggesting that *DPP4* blocks the fibroblast growth factor signaling pathway. A prostate tumor cell line (*1-LN*) has been reported to be invasion resistant by blocking plasminogen binding to *DPP4* ***(52)***.

SPP1/OPN (*osteopontin*) is an integrin-binding glycoprotein of the extracellular matrix with many functions ***(53)***. It is a proven mediator of tumorigenesis in several cancers, and has previously been proposed to be a potential predictor of malignancy in prostate tumors ***(54,55)***. The highly malignant carcinoma tissue had an *SPP1/OPN* increase of up to sixfold in comparison with normal tissue. Strong *OPN* expression was observed in the normal human endometrium in 80% of the samples analyzed in ***(56)***; in endometriod carcinoma *SPP1/OPN* expression levels were low or not observed, whereas serous tumors displayed over-expression. It appears that *SPP1/OPN* enhances cell proliferation ***(57)*** via the epidermal growth factor pathway, at least in the LNCaP prostate cancer cell line ***(58)***.

2.5. Analysis of Differentially Expressed Genes and Pathways

2.5.1. PDGF Signaling and Gleason Pattern 3

Currently, the relationship of *MAOA* and prostate cancer is not clear. However, in our signature we have also identified *PDGFB (Platelet-derived growth factor β polypeptide (simian sarcoma viral (v-sis) oncogene homolog)* as part of our genetic signature. This is important since we have identified the Platelet-Derived Growth Factor Signaling Pathway as the most significantly differentiated pathway (using as input the genes that best discriminate between samples with Gleason pattern 3 and other types). Using Bonferroni correction for multiple testing from the Panther Classification Gene Expression tool, the PDGF Signaling Pathway has a P-value lower than 0.000176 (see **Fig. 4**). In the molecular signatures of Gleason-3-versus-(4+5), Gleason-4-versus-(3+5) and Gleason-5-versus-(3+4), we have identified a number of genes differentially expressed in this pathway. We highlight *STAT1* and *STAT6 (signal transducer and activator of transcription 6, interleukin-4 induced)*. *STAT6* has not been identified previously ***(4)***, but it is well-correlated across all samples with *IL-4 (interleukin 4, M13982)*, *IRF4 (Interferon regulatory factor 4)* and *GSTM1 (glutathione S-transferase M1)* which is part of the authors' 86 most discriminatory gene set. While *STAT6* is up-regulated in most Gleason 3 samples, *STAT1* is down-regulated. *STAT6* was previously implicated in prostate cancer by Ni and colleagues ***(59)*** where significant levels of activated *STAT6* and *STAT4* were detected in primary prostate tissues but no significant expression of active *STAT1*, *STAT2*, or *STAT5* had been detected. A novel statistical method used to integrate microarray data also points at *STAT6* involvement in prostate cancer ***(60)***. *STAT6* (CR606877) is also highly correlated with the expression of *PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1 (p85 alpha), CR977491*), but the latter has a non-uniform expression in Gleason 3 samples and might not be useful as a biomarker.

Other genes in the PDGF signaling pathway, which are differentially expressed in Gleason 3 samples in comparison with other sample types include *PIK3C3* (*phosphoinositide-3-kinase, class 3*), *ARHGAP4 (Rho GTPase activating protein 4,CD359532)*, *USF2 (upstream transcription factor 2, c-fos interacting, S50537)*, and *MAPK3 (mitogen-activated protein kinase 3)*. Interestingly, *MAPK3* and *USF2* appear to co-express across the whole sample set (both Gleason 3 and those that are not), while *ARHGAP4 c*orrelates well with *MYC (V-myc myelocytomatosis viral oncogene homolog (avian) BT019768)* in almost all samples but not those with a Gleason pattern 3. *ARHGAP4* appears to have a conspicuously similar expression value in normal and cancer tissue, while *MYC* shows marked differences (in Gleason 3 samples).

2.5.2. Other Pathways

There are other pathways that appear to be having a significantly higher number of genes differentially expressed in them, including *integrin signaling, inflammation mediated by chemokine and cytokine signaling, androgen/estrogen/progesterone biosynthesis, p53 pathway feedback loops 2,* and *endothelin signaling.* The genetic signature of Gleason-4-versus-(3+5) seems to indicate that Gleason 4 appears to be linked to *inflammation* (*P*-value < 1.42E-02) and Gleason-5-versus-(3+4) points to *integrin signaling* (*P*-value < 1.42E-02).

2.5.3. Pathway Analysis from the Union of all the Signatures

Aside from individual gene lists where each one is associated with one of the patterns, we can have a single list that represents the union of all genes present in at least one of the molecular signatures. Mapping them to pathways should provide a better understanding of the most significant patterns involved in prostate cancer that is less biased by individual subtypes. This approach has revealed *FGF signaling, oxidative stress response, RAS, B cell activation, hypoxia response via HIF activation,* and *VEGF signaling* as putative pathways of interest for a more comprehensive genome-wide study (**Table 1**).

2.5.4. A Critical Analysis of the Labeling Produces more Coherent Molecular Signatures

We have identified four samples that do not seem to have a profile that is similar to the ones that have been categorized using the same labeling in our signatures. They are 03-060A, 02-003E, and 03-115E (that appear to have a Gleason pattern 4 profile) and 03-135C (which exhibits an apparent Gleason pattern 5 profile). This is relatively clear from detailed inspection of **Fig. 2** (we discussed how we reached these conclusions in the caption for this figure).

With these modifications, we have computed new molecular signatures. As a consequence, the Gleason-3-versus-(4+5) is now the (289,254)-571-feature set (**Fig. 3a**) and is a more coherent signature for this pattern with a slight increase in the total number of probes used (from 522 to 571). Analogously, molecular signatures for Gleason-4-versus-(3+5) and Gleason-5-versus-(3+4) correspond to (114,124)-243 and (165,151)-262 feature sets respectively (**Figs. 3b** and **3c**, respectively).

Table 1
The pathways differentially expressed identified using the union of all the genes present in at least one of the molecular signatures for Gleason patterns 3, 4 and 5 (see Fig. 2). The first column (total) indicates the number of genes in the pathway; the second (labelled #genes) indicates the number of genes of that pathway present in the union of the molecular signature. The third column (# expected) indicates the expected number of genes to appear in the pathway and finally a P-value is associated to the pathway taking into account Bonferroni correction for multiple testing. Data generated using a public domain tool available from (http://www.pantherdb.org) *(61)*.

Pathway	Total	# genes	# expected	*P*-value
Integrin signaling pathway	236	20	4.19	2.07E-06
PDGF signaling pathway	180	17	3.19	5.46E-06
Inflammation	314	22	5.57	1.08E-05
p53 pathway feedback loops 2	65	8	1.15	3.48E-03
EGF receptor signaling pathway	136	11	2.41	5.43E-03
Angiogenesis	219	14	3.88	6.55E-03
Endothelin signaling pathway	98	9	1.74	1.05E-02
Androgen/estrogen/progesterone biosynthesis	18	4	0.32	4.34E-02
p53 pathway	120	9	2.13	4.62E-02
T cell activation	120	9	2.13	4.62E-02
Huntington disease	177	11	3.14	5.20E-02
FGF signaling pathway	135	9	2.39	1.06E-01
Oxidative stress response	67	6	1.19	1.83E-01
Ras Pathway	92	7	1.63	1.90E-01
B cell activation	99	7	1.76	2.85E-01
Hypoxia response via HIF activation	33	4	0.59	3.99E-01
VEGF signaling pathway	79	6	1.4	4.12E-01

3. Conclusions

Overall, there seems to be very distinct phenotypic types of tumors within the histopathologically classified group labelled Gleason pattern 4 as well as in Gleason pattern 5. Nevertheless, we have been able to identify similarities between these Gleason patterns where other methods have failed. This is now more evident from the inspection of the profiles of samples for the gene subset composed of *CRABP2, TPM2, EDNRA, CTGF, CXCR4, DPP4, SPP1, MAOA, DAD1, KL6, MYBPC1, SPON2* (as well as other closely co-expressed genes

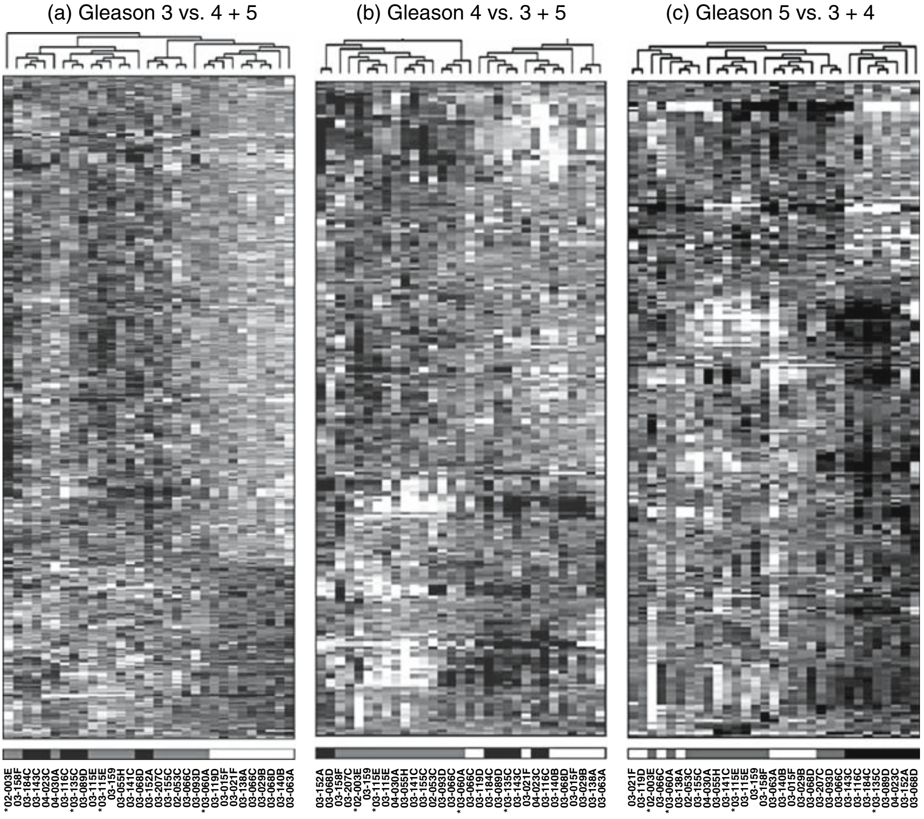


Fig. 3. Molecular signatures for Gleason 3, 4, and 5, in that order from left to right. After modifying the labels of the four samples, which in the previous figure appear to belong to different groups, new molecular signatures for the three groups can be observed. This has produced a more coherent molecular signature for the groups Gleason pattern 3, 4, and 5 (now slightly modified).

like *CABIN1*, to be discussed below), can lead to a different classification of prostate tumors. Validation of this molecular taxonomy with immunohistochemical methods and RT-PCR would be required for the application and study of its relevance in the clinical setting, both necessary steps for translational medical research.

In a study where a number of genes associated with aggressiveness in androgen-independent metastatic tumors were up-regulated, the putative tumor suppressor gene *KLF6* ***(62**,**63)*** was decreased ***(64)***. The over-expression of *MYBPC1* in Gleason 3 is a significant finding since *MYBPC1* has been

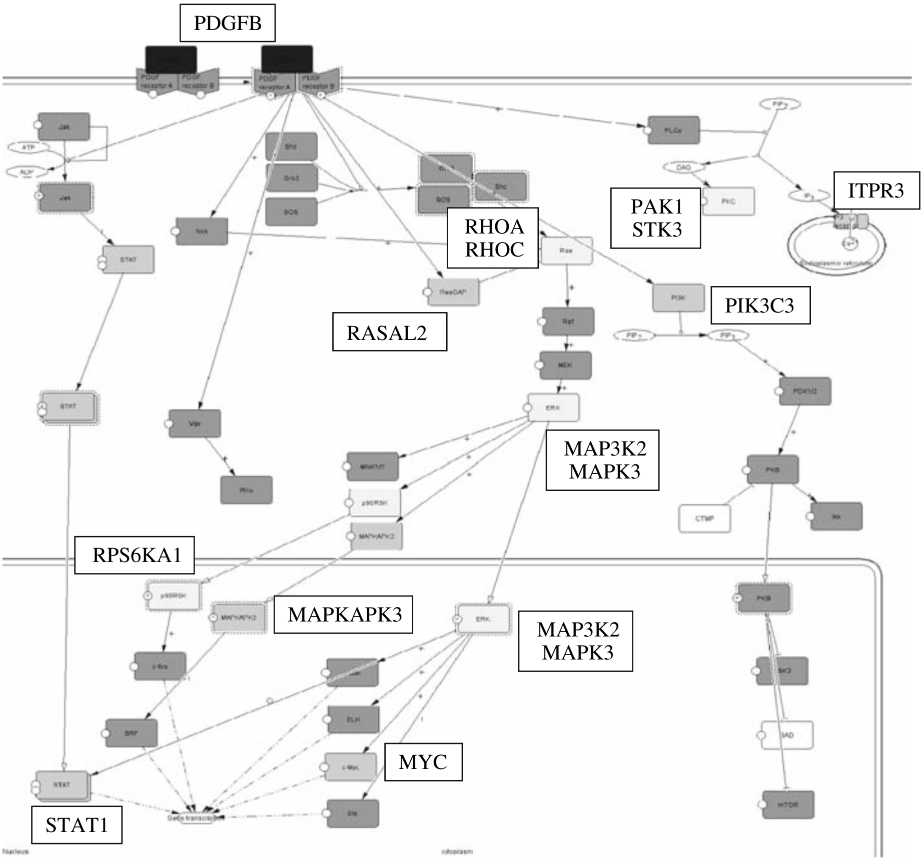


Fig. 4. *Platelet-derived growth factor signaling* (above) is one of the most discriminated pathways for Gleason pattern 3 in our analysis (*P*-value < 8.12E-03). In gray we highlight genes (e.g., *RasGAP*, *STAT*) which are differentially expressed in the signature for Gleason-3-versus-(4+5) (in light gray if they appear in more than one signature of **Fig. 3**, e.g., *PKC*, *Ras*, etc). After our change of labeling of four samples and subsequent reanalysis, *inflammation mediated by cytokines and chemokines* and *integrin signaling* are still the most statistically significant pathways for Gleason patterns 4 and 5, respectively. However, the *RAS pathway* is now the most discriminatory one for Gleason pattern 3 with a *P*-value < 7.66E-03, and following *PDGF signaling* we have now have for Gleason 3: *T-cell activation* (*P*-value < 4.58E-02), *interleukin signaling* and *endothelin signaling* (both with *P*-values < 7.13E-02).

implicated in severe myopathies and in laryngeal squamous cell carcinoma ***(65)***. Recent analyses have shown that it is a short-lived proteasomal substrate; and that over expression of *USPm* (the longer ubiquitin-specific protease isoform) prevents its degradation ***(66)***. This has opened a number of interesting scenarios with new working hypotheses, which are starting to be explored. Those include *USP25*, which has an essential role in protein degradation *via* the 26S proteasome and thus regulates several cellular pathways. The down-regulation of *CABIN1 (calcineurin binding protein 1)* in the samples which are not Gleason 3*,* (and up-regulation in our one-sample-modified Gleason 3) is also worth of note. Calcineurin is a phosphoprotein phosphatase that channels intracellular calcium signals into several biological pathways ***(67)***. Calcineurin-NFAT signaling has a critical role in T-cell activation and *CABIN1* plays a major role as transcriptional co-repressor of *myocyte enhancer 2* (*MEF2*) ***(68,69)***.

We have presented the application of a number of new mathematical models and algorithms developed during the past years that allows a molecular classification of subtypes of a given disease *via* interrogation of gene expression profiles. The methods involve the categorization of numerical data for a posterior analysis with combinatorial optimization methods for gene selection. We have proven the adequacy of the tandem of methodologies in the very difficult problem of finding biomarkers of interest in prostate cancer. Our approach challenges the clinical identification of tumor subtypes and as a consequence may have a great impact on translational studies. It is also clear to us that current large-scale clinical trials and studies require similar tools to allow them to deal with the unquestionable intrinsic diversity and similarity of the human transcriptome in both health and disease.

4. Note

1. Although the Gleason system for grading prostate cancer has withstood the test of time, it can still be associated with discrepancies. The most important is sampling error since only small amounts of tissue are removed with the thin needle core biopsies. Because the grading is based entirely on the histologic pattern of the tumor cells in stained sections, another possible source of error is the experience of the person examining the tissue and making the report. Finally, it should be noted that prostate cancer is a complex disease and the Gleason score alone cannot predict the outcome, e.g., some with low scores end up with a poor clinical result and vice versa. Hence, the continued interest in enhancing the Gleason grading with additional and more objective parameters such as gene expression via microarray.

References

1. Cotta, C., Sloper, C., and Moscato, P. (2004) Evolutionary search of thresholds for robust feature set selection: application to the analysis of microarray data, in *Applications of Evolutionary Computing*, (Raidl, G., et al., eds.), *Lecture Notes in Computer Science* 3005, Springer-Verlag, Berlin, pp. 21–30.
2. Moscato, P., Mendes, A., and Berretta, R. (2007) Benchmarking a memetic algorithm for ordering microarray data, *Biosystems* **88**(1–2), 56–75.
3. Gleason, D. F. (1966) Classification of prostatic carcinomas. *Cancer Chemother. Rep*. **50**, 125–128.
4. True, L., Coleman, I., Hawley, S., Huang, C. -Y., Gifford, D., Coleman, R., et al. (2006) A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proc. Nat. Acad. Sci. U S A* **103**(29), 10991–10996.
5. Dehm, S. M., and Tindall, D. J. (2006) Molecular regulation of androgen action in prostate cancer. *J. Cell Biochem.* **99**, 333–344.
6. Chan, J. M., Stampfer, M. J., Giovannucci, E., Gann, P. H., Ma, J., Wilkinson, P., et al. (1998) Plasma insulin-like growth factor-1 and prostate cancer risk: a prospective study. *Science* **279**, 563–566.
7. Pollack, M. N., Schernhammer, E. S., and Hankinson, S. E. (2004) Insulin-like growth factor and neoplasia. *Nat. Rev. Cancer* **4**, 505–518.
8. Wu, J. D., Haugk, K., Woodke, L., Nelson, P., Coleman, I., and Plymate, S. R. (2006) Interaction of IGF signaling and the androgen receptor in prostate cancer progression. *J. Cell Biochem.* **99**, 392–401.
9. He, W. W., Sciavolino, P. J., Wing, J., Augustus, M., Hudson, P., Meissner, P. S., et al. (1997) A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* **43**, 69–77.
10. Myers, M. P., Pass, I., Batty, I. H., Van der Kaay, J., Stolarov, J. P., Hemmings, B. A., et al. (1998) The lipid phosphatase activity of PTEN is critical for its tumor suppressor function. *Proc. Natl. Acad. Sci U S A* **95**, 13515–13518.
11. Di Cristofano, A., De Acetis, M., Koff, A., Cordon-Cardo, C., and Pandolfi, P. P. (2001) PTEN and p27Kip1 cooperate in prostate cancer tumour suppression in the mouse. *Nat. Genet*. **27**, 222–224.
12. Kaspar, P., Dvorakova, M., Kralova, J., Pajer, P., Kozmik, Z., and Dvorak, M. (1999) Myb-interacting protein, ATBF1, represses transcriptional activity of Myb oncoprotein. *J. Biol. Chem*. **274**, 14422–14428.
13. Narla, G., Heath, K. E., Reeves, H. L., Li, D., Giono, L. E., Kimmelman, A. C., et al. (2001) KLF6, a candidate tumor suppressor gene mutated in prostate cancer. *Science* **294**, 2563–2566.
14. Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648.

15. Berretta, R., Costa, W., and Moscato, P. (2006) Combinatorial optimization models for finding genetic signatures from gene expression datasets, in *Bioinformatics—Methods in Molecular Biology Series* (Keith, J., ed.), Humana Press, Totowa, NJ.
16. Houran, M., Berretta, R., Mendes, A., and Moscato, P. (2006) Genetic signatures for a Rodent model of Parkinson's disease using combinatorial optimization methods, in *Bioinformatics—Methods in Molecular Biology Series* (Keith, J., ed.), Humana Press, Totowa NJ.
17. Mahata, P., Costa, W., Cotta, C., and Moscato, P. (2006) Hierarchical clustering, languages and cancer. *Proceedings of EvoBIO2006—4th European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics* , in: *Lecture Notes in Computer Science* **3907** (Rothlauf, F., et al., eds.), Budapest, Hungary, pp. 67–78.
18. Moscato, P., and Cotta, C. (2006) Memetic algorithms, in *Handbook of Approximation Algorithms and Metaheuristics* (Gonzalez, T. F., ed.), Chapman & Hall/CRC, London.
19. Cotta, C., Langston, M., and Moscato, P. (2006) Combinatorial and algorithmic issues for microarray data analysis, in *Handbook of Approximation Algorithms and Metaheuristics* (Gonzalez, T. F., ed.), Chapman & Hall/CRC, London.
20. Edwards, S., Campbell, C., Flohr, P., Shipley, J., Giddings, I., Te-Poele, R., et al. (2005) Expression analysis onto microarrays of randomly selected cDNA clones highlights HOXB13 as a marker of human prostate cancer. *Br. J. Cancer* **92**, 376–381.
21. Parry, R., Schneider, D., Hudson, D., Parkes, D., Xuan, J. A., Newton, A., et al. (2005) Identification of a novel prostate tumor target, mindin/RG-1, for antibody-based radiotherapy of prostate cancer. *Cancer Res.* **65**, 8397–3405.
22. Okuducu, A. F., Janzen, V., Ko, Y., Hahne, J. C., Lu, H., Ma, Z. L., et al. (2005) Cellular retinoic acid-binding protein 2 is down-regulated in prostate cancer. *Int. J. Oncol.* **27**, 1273–1282.
23. Donato, L. J., and Noy, N. (2005) Suppression of mammary carcinoma growth by retinoic acid: proapoptotic genes are targets for retinoic acid receptor and cellular retinoic acid-binding protein II signalling. *Cancer Res.* **65**, 8193–8199.
24. Budhu, A. S., and Noy, N. (2002) Direct channeling of retinoic acid between cellular retinoic acid-binding protein II and retinoic acid receptor sensitizes mammary carcinoma cells to retinoic acid-induced growth arrest. *Mol. Cell. Biol.* **22**, 2632–2641.
25. Yang, F., Tuxhorn, J. A., Ressler, S. J., McAlhany, S. J., Dang, T. D., and Rowley, D. R. (2005) Stromal expression of connective tissue growth factor promotes angiogenesis and prostate cancer tumorigenesis. *Cancer Res.* **65**, 8887–8895.
26. Suzuki, K., Obara, K., Kobayashi, K., Yamana, K., Bilim, V., Itoi, T., et al. (2006) Role of connective tissue growth factor in fibronectin synthesis in cultured human prostate stromal cells. *Urology* **67**, 647–653.

27. Cazaubon, S., Deshayes, F., Couraud, P. O., and Nahmias, C. (2006) Endothelin-1, angiotensin II and cancer. *Med. Sci. (Paris)* **22**, 416–422.
28. Herrmann, E., Bogemann, M., Bierer, S., Eltze, E., Hertle, L., and Wulfing, C. (2006) The endothelin axis in urologic tumors: mechanisms of tumor biology and therapeutic implications. *Expert Rev. Anticancer Ther.* **6**, 73–81.
29. Nelson, J. B., Udan, M. S., Guruli, G., and Pflug, B. R. (2005) Endothelin-1 inhibits apoptosis in prostate cancer. *Neoplasia* **7**, 631–637.
30. Bagnato, A., and Natali, P. G. (2004) Endothelin receptors as novel targets in tumor therapy. *J. Transl. Med.* **2**, 16.
31. Varga, A. E., Stourman, N. V., Zheng, Q., Safina, A. F., Quan, L., Li, X., et al. (2005) Silencing of the Tropomyosin-1 gene by DNA methylation alters tumor suppressor function of TGF-beta. *Oncogene* **24**, 5043–5052.
32. Wang, J., Wang, J., Sun, Y., Song, W., Nor, J. E., Wang, C. Y., et al. (2005) Diverse signaling pathways through the SDF-1/CXCR4 chemokine axis in prostate cancer cell lines leads to altered patterns of cytokine secretion and angiogenesis. *Cell Signal.* **17**, 1578–1592.
33. Taichman, R. S., Cooper, C., Keller, E. T., Pienta, K. J., Taichman, N. S., and McCauley, L. K. (2002) Use of the stromal cell-derived factor-1/CXCR4 pathway in prostate cancer metastasis to bone. *Cancer Res.* **62**, 1832–1837.
34. Darash-Yahana, M., Pikarsky, E., Abramovitch, R., Zeira, E., Pal, B., Karplus, R., et al. (2004) Role of high expression levels of CXCR4 in tumor growth, vascularization, and metastasis. *FASEB J.* **18**, 1240–1242.
35. Mochizuki, H., Matsubara, A., Teishima, J., Mutaguchi, K., Yasumoto, H., Dahiya, R., et al. (2004) Interaction of ligand-receptor system between stromal-cell-derived factor-1 and CXC chemokine receptor 4 in human prostate cancer: a possible predictor of metastasis. *Biochem. Biophys. Res. Commun.* **320**, 656–663.
36. Singh, S., Singh, U. P., Grizzle, W. E., and Lillard, J. W. Jr. (2004) CXCL12–CXCR4 interactions modulate prostate cancer cell migration, metalloproteinase expression and invasion. *Lab. Invest.* **84**, 1666–1676.
37. Arya, M., Patel, H. R., McGurk, C., Tatoud, R., Klocker, H., Masters, J., et al. (2004) The importance of the CXCL12–CXCR4 chemokine ligand–receptor interaction in prostate cancer metastasis. *J. Exp. Ther. Oncol.* **4**, 291–303.
38. Li, S., Huang, S., and Peng, S. B. (2005) Overexpression of G protein-coupled receptors in cancer cells: involvement in tumor progression. *Int. J. Oncol.* **27**, 1329–1339.
39. Hart, C. A., Brown, M., Bagley, S., Sharrard, M., and Clarke, N. W (2005) Invasive characteristics of human prostatic epithelial cells: understanding the metastatic process. *Br. J. Cancer* **92**, 503–512.
40. Engl, T., Relja, B., Marian, D., Blumenberg, C., Muller, I., Beecken, W. D., et al. (2006a) CXCR4 chemokine receptor mediates prostate tumor cell adhesion through alpha5 and beta3 integrins. *Neoplasia* **8**, 290–301.

41. Berquin, I. M., Min, Y., Wu, R., Wu, H., and Chen, Y. Q. (2005) Expression signature of the mouse prostate. *J. Biol. Chem.* **280**, 36442–36451.
42. Chinni, S. R., Sivalogan, S., Dong, Z., Filho, J. C., Deng, X., Bonfil, R. D., et al. (2006) CXCL12/CXCR4 signaling activates Akt-1 and MMP-9 expression in prostate cancer cells: the role of bone microenvironment-associated CXCL12. *Prostate* **66**, 32–48.
43. Havens, A. M., Jung, Y., Sun, Y. X., Wang, J., Shah, R. B., Buhring, H. J., et al. (2006) The role of sialomucin CD164 (MGC-24v or endolyn) in prostate cancer metastasis. *BMC Cancer* **6**, 195.
44. Vaday, G. G., Hua, S. B., Peehl, D. M., Pauling, M. H., Lin, Y. H., Zhu, L., et al. (2004) CXCR4 and CXCL12 (SDF-1) in prostate cancer: inhibitory effects of human single chain Fv antibodies. *Clin. Cancer Res.* **10**, 5630–5639.
45. Sun, Y. X., Schneider, A., Jung, Y., Wang, J., Dai, J., Wang, J., et al. (2005) Skeletal localization and neutralization of the SDF-1(CXCL12)/CXCR4 axis blocks prostate cancer metastasis and growth in osseous sites *in vivo*. *J. Bone Miner. Res.* **20**, 318–329.
46. Miwa, S., Mizokami, A., Keller, E. T., Taichman, R., Zhang, J., and Namiki, M. (2005) The bisphosphonate YM529 inhibits osteolytic and osteoblastic changes and CXCR-4-induced invasion in prostate cancer. *Cancer Res.* **65**, 8818–8825.
47. Kukreja, P., Abdel-Mageed, A. B., Mondal, D., Liu, K., and Agrawal, K. C. (2005) Up-regulation of CXCR4 expression in PC-3 cells by stromal-derived factor-1alpha (CXCL12) increases endothelial adhesion and transendothelial migration: role of MEK/ERK signaling pathway-dependent NF-kappaB activation. *Cancer Res.* **65**, 9891–9898.
48. Engl, T., Relja, B., Blumenberg, C., Muller, I., Ringel, E. M., Beecken, W. D., et al. (2006b) Prostate tumor CXC-chemokine profile correlates with cell adhesion to endothelium and extracellular matrix. *Life Sci.* **78**, 1784–1793.
49. Oudes, A. J., Campbell, D. S., Sorensen, C. M., Walashek, L. S., True, L. D., and Liu, A. Y. (2006) Transcriptomes of human prostate cells. *BMC Genomics* **7**, 92.
50. Wilson, M. J., Ruhland, A. R., Quast, B. J., Reddy, P. K., Ewing, S. L., and Sinha, A. A. (2000) Dipeptidylpeptidase IV activities are elevated in prostate cancers and adjacent benign hyperplastic glands. *J. Androl.* **21**, 220–226.
51. Wesley, U. V., McGroarty, M., and Homoyouni, A. (2005) Dipeptidyl peptidase inhibits malignant phenotype of prostate cancer cells by blocking basic fibroblast growth factor signaling pathway. *Cancer Res.* **65**, 1325–1334.
52. Gonzalez-Gronow, M., Grenett, H. E., Gawdi, G., and Pizzo, S. V. (2005) Angiostatin directly inhibits human prostate tumor cell invasion by blocking plasminogen binding to its cellular receptor, CD26. *Exp. Cell. Res.* **303**, 22–31.
53. Chabas, D. (2005) Osteopontin, a multi-faceted molecule. *Med. Sci. (Paris)* **21**, 832–838.

54. Tozawa, K., Yamada, Y., Kawai, N., Okamura, T., Ueda, K., and Kohri, K. (1999) Osteopontin expression in prostate cancer and benign prostatic hyperplasia. *Urol. Int.* **62**, 155–158.
55. Forootan, S. S., Foster, C. S., Aachi, V. R., Adamson, J., Smith, P. H., Lin, K., et al. (2006) Prognostic significance of osteopontin expression in human prostate cancer. *Int. J. Cancer* **118**, 2255–2261.
56. Briese, J., Schulte, H. M., Bamberger, C. M., Loning, T., and Bamberger, A. M. (2006) Expression pattern of osteopontin in endometrial carcinoma: correlation with expression of the adhesion molecule CEACAM1. *Int. J. Gynecol. Pathol.* **25**, 161–169.
57. Elgavish, A., Prince, C., Chang, P. L., Lloyd, K., Lindsey, R., and Reed, R. (1998) Osteopontin stimulates a subpopulation of quiescent human prostate epithelial cells with high proliferative potential to divide *in vitro*. *Prostate* **35**, 83–94.
58. Angelucci, A., Festuccia, C., Gravina, G. L., Muzi, P., Bonghi, L., Vicentini, C., et al. (2004) Osteopontin enhances the cell proliferation induced by the epidermal growth factor in human prostate cancer cells. *Prostate* **59**, 157–166.
59. Ni, Z., Lou, W., Lee, S. O., Dhir, R., DeMiguel, F., Grandis, J. R., et al. (2002) Selective activation of members of the signal transducers and activators of transcription family in prostate carcinoma. *J. Urol.* **167**, 1859–1862.
60. Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., and Winslow, R. L. (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics* **21**, 3905–3911.
61. Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., et al. (2006) B. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* **34**, W645–W650.
62. Narla, G., Difeo, A., Reeves, H. L., Schaid, D. J., Hirshfeld, J., Hod, E., et al. (2005) A germline DNA polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk. *Cancer Res.* **65**, 1213–1222.
63. Rubinstein, M., Idelman, G., Plymate, S. R., Narla, G., Friedman, S. L., and Werner, H. (2004) Transcriptional activation of the insulin-like growth factor I receptor gene by the Kruppel-like factor 6 (KLF6) tumor suppressor protein: potential interactions between KLF6 and p53. *Endocrinology* **145**, 3769–3777.
64. Narla G, Heath K. E., Reeves H. L., Li D., Giono L. E., Kimmelman A. C., et al. (2005) Targeted inhibition of the KLF6 splice variant, KLF6 SV1, suppresses prostate cancer cell growth and spread. *Cancer Res.* 65:5761–5768.
65. Beier, U. H., Gorogh, T., Holtmeier, C., Ambrosch, P., and Maune, S. (2002) Overexpression of the human myosin-binding protein-C1 mRNA in laryngeal squamous cell carcinoma cells. *Anticancer Res.* **22**, 3343–3347.

66. Bosch-Comas, A., Lindsten, K., Gonzalez-Duarte, R., Masucci, M. G., and Marfany, G. (2006) The ubiquitin-specific protease USP25 interacts with three sarcomeric proteins. *Cell Mol. Life Sci.* **63**, 723–734.
67. Mignen, O., Thompson, J. L., and Shuttleworth, T. J. (2003) Calcineurin directs the reciprocal regulation of calcium entry pathways in nonexcitable cells. *J. Biol. Chem.* **278**, 40088–40096.
68. Youn, H. D., and Liu, J. O. (2000) Cabin1 represses MEF2-dependent Nur77 expression and T cell apoptosis by controlling association of histone deacetylases and acetylases with MEF2. *Immunity* **13**, 85–94.
69. Pan, F., Means, A. R., and Liu, J. O. (2005) Calmodulin-dependent protein kinase IV regulates nuclear export of Cabin1 during T-cell activation. *EMBO J.* **24**, 2104–2113.

9

Microarrays for the Study of Viral Gene Expression During Human Cytomegalovirus Latent Infection

Barry Slobedman and Allen K. L. Cheung

Summary

Human cytomegalovirus (HCMV) is one of the largest known DNA viruses. It is ubiquitous, and following resolution of primary productive infection, it persists in the human host by establishing a lifelong latent infection in myeloid lineage cells such as monocytes and their progenitors. Most adults with HCMV infection are healthy but it can cause neurologic deficits in infants, and remains an important cause of morbidity and mortality in the immunosuppressed patient. Microarray-based studies of HCMV have provided useful information about genes that are transcriptionally active during both productive and latent phases of infection. This chapter describes how to study genes in HCMV using microarrays and two cell types (productively infected human foreskin fibroblasts, and latently infected primary human myeloid progenitor cells).

Key Words: microarray, herpesvirus human cytomegalovirus latent infection, myeloid progenitor cell, viral gene transcription.

Abbreviations: aRNA – amplified RNA; HCMV – human cytomegalovirus; HFF – human foreskin fibroblast; ORF – open reading frame; RT-PCR – reverse transcriptase PCR

1. Introduction

Human cytomegalovirus (HCMV) is a species-specific β-herpesvirus and is one of the largest known DNA viruses, with a double-stranded linear genome of approximately 230 kb ***(1)***. The genome consists of a long (L) and a short (S) segment, each of which is flanked by inverted repeat sequences. The prototype

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

HCMV strain AD169 has been fully sequenced, and its original annotation identified no fewer than 208 potential open reading frames (ORFs) ***(2)***. Other laboratory strains of HCMV, such as the Towne strain, have approximately 5 kb of DNA sequence not found in AD169, and the low passage strain, Toledo, contains 15 kb of sequence not found in either AD169 or Towne strains. The extra sequences present in the Towne and Toledo strains are predicted to encode 4 and 19 additional ORFs, respectively ***(3)***. In recent years, the protein-coding potential across these strains has been reassessed, with current estimates ranging from 165–192 protein-encoding ORFs, depending on the criteria used for sequence analysis ***(4,5)***. Additional analyses including a broader range of clinical HCMV clinical isolates have identified a total of no fewer than 252 potential protein encoding ORFs ***(6)***.

HCMV is ubiquitous, with infection rates approaching 90% in some communities. Following immune-mediated resolution of a primary productive HCMV infection, the virus is able to persist in the host by establishing a lifelong latent infection that is not cleared by the immune response ***(1,7,8)***. Latency is characterized by maintenance of the viral genome in the absence of infectious virus production and with restricted viral gene expression, although the full repertoire of genes expressed during the latent phase has not been defined. Periodically, virus may reactivate from latency, resulting in the production of new, infectious virus. The virus is extremely well adapted to its host with the majority of productive HCMV infections being mild or asymptomatic in healthy adults. However, HCMV is the most common congenitally acquired infection in infants where it is the leading viral cause of neurological defects ***(9–12)***. Reactivation of virus from latency is a major clinical concern to those undergoing immunosuppressive therapies such as bone marrow and solid organ allogeneic transplant recipients, where serious, frequently life-threatening HCMV-associated disease is common during the post-transplant period ***(12)***.

HCMV DNA can be detected in myeloid cells from peripheral blood of healthy seropositive individuals in the absence of detectable infectious virus ***(13–15)***. In addition, virus has been reactivated from terminally differentiated monocyte-derived macrophages ***(16)*** or differentiated dendritic cell precursors ***(17)***, implicating myeloid lineage cells as an important site of latent infection. Despite the importance of the latent phase of infection to the success of this virus as a human pathogen, little is known about viral gene expression during the establishment and maintenance of latency, and the global assessment of HCMV gene expression has been complicated by the large number of potential viral ORFs that may be expressed.

The use of microarray technologies has provided a unique opportunity to screen rapidly large viruses such as herpesviruses for transcriptional activity during infection of a variety of cell types ***(18–24)***. The procedure can be divided into five main steps: (1) amplification of viral gene sequences; (2) generation of viral gene microarrays bearing almost all known HCMV genes; (3) extraction, amplification, and labeling of cDNAs; (4) detection of global HCMV gene transcription following hybridization of labeled infected cell cDNAs to viral gene microarrays; (5) confirmation of viral gene expression by reverse transcription PCR (RT-PCR).

2. Materials

2.1. Amplification of Viral Gene Sequences

1. HCMV gene-specific PCR primers
2. Cell lysis buffer ***(25)***: 50 mM KCl (ICN Biomedicals, Aurora, OH), 10 mM Tris-HCl pH 8.5, 2 mM $MgCl_2$, 0.45% Nonidet P-40, 0.45% Tween 20, Proteinase K (100 μg/mL) (Invitrogen, Carisbad, CA).
3. PCR Supermix (Invitrogen)
4. MultiScreen PCR_{96} filter plates (Millipore, Billerica, MA)
5. MultiScreen vacuum manifold (Millipore)

2.2. Purification and Spotting of Viral Gene Sequences onto Glass Slides

1. Poly-L-lysine coated glass microscope slides
2. Microarrayer Robot (ESI Inc, Toronto, Ontario, Canada)
3. SSC (20×, Sigma)
4. SDS (10% (w/v), GIBCO, Invitrogen)
5. UV Stratalinker 1800 (Stratagene, La Jolla, CA)
6. Succinic anhydride (Sigma)
7. 1-methyl-2-pyrrolidinone (Sigma)
8. 1M sodium borate (pH 8.0, filtered; Sigma)
9. MWG Spotting Buffer (MWG Biotech Inc, Highpoint, NC)

2.3. RNA Extraction, Amplification, and Labeling of cDNAs

1. RNAqueous kit, 50 purifications (Ambion, Austin, TX)

   a. Lysis/Binding Solution
   b. 64% ethanol
   c. Wash Solution #1
   d. Wash Solution #2/3
   e. Elution Tubes
   f. Filter Cartridge

2. 5M Ammonium acetate (Ambion)
3. 0.1 μg/μL linear acrylamide (Ambion)
4. MessageAmp II aRNA amplification kit (Ambion):

    a. T7 Oligo(dT) primer
    b. 10× First Strand Buffer
    c. dNTP Mix
    d. RNase inhibitor
    e. ArrayScript
    f. 10× Second Strand Buffer
    g. DNA Polymerase
    h. RNase H
    i. Binding Buffer
    j. 10X Reaction Buffer
    k. T7 NTP solution
    l. T7 Enzyme Mix
    m. cDNA Elution Tube + Filter Cartridge
    n. aRNA Filter Cartridge
    o. Collection Tube

5. Reverse Transcription and Labeling of cDNA

    a. Oligo-dT primers (3 μg/μL stock, Invitrogen)
    b. Random primers (3 μg/μL stock, Invitrogen)
    c. 5× First Strand Buffer (supplied with Superscript II, Invitrogen)
    d. 0.1M DTT (supplied with Superscript II, Invitrogen)
    e. Un-labeled dNTP mix (10 mM stock, Invitrogen)
    f. Superscript II (200 U/μL stock, Invitrogen)
    g. Cya-3-dUTP (0.1 mM, Promega, Madison, WI)
    h. Cya-5-dUTP (0.1 mM, Promega)
    i. NaOH (0.1N, Sigma)
    j. HCl (0.1N, Sigma)
    k. TE buffer (100× stock; 10 mM Tris-HCl, 1 mM EDTA, pH~8.0, 0.2 μm-filtered, Sigma)

6. Microcon YM-30 filters (Millipore)
7. Cot-1 human DNA (1mg/mL, Invitrogen)
8. poly-A RNA (10mg/mL, Invitrogen)
9. tRNA (10mg/ml, Invitrogen)

2.4. Microarray Hybridization and Washing

1. TE buffer
2. 20× SSC
3. 10% (w/v) SDS

4. 22 × 50-mm glass coverslips (Menzel Glaser GmbH, Fisher Scientific, Braunschweig, Germany)
5. Glass Array Hybridization Cassette (Ambion)
6. 350-ml glass slide chambers (Shandon Lipshaw, Pittsburgh, PA)
7. Slide rack (holds 30 slides) (Shandon Lipshaw)

2.5. Validation of Microarray Results using RT-PCR

1. RQ1 DNase (Promega)
2. RQ1 Buffer (Promega)
3. RQ1 Stop buffer (Promega)
4. Superscript II reverse transcriptase (Invitrogen)—includes 5× First Strand Buffer, 0.1M DTT
5. 25 mM dNTPs mix (dATP, dTTP, dGTP, dCTP, Invitrogen)
6. Random Primers (3 mg/mL stock, Invitrogen)
7. RNaseOUT (40 U/μL, Invitrogen)
8. PCR Platinum Taq (Invitrogen)—includes 10× PCR Buffer, 2 mM $MgCl_2$
9. Gene specific primers (Proligos, Sigma)

3. Methods

3.1. Amplification of Viral Gene Sequences

1. Viral genomic DNA template for subsequent PCR amplification of individual viral gene sequences was generated from human foreskin fibroblast (HFFs) infected with 3 strains of HCMV—AD169, Toledo, and Towne (*see* **Note 1**)
2. Harvest and wash cells three times in PBS
3. Add 500 μL of Cell Lysis Buffer to 5 × 10^6 washed cells and overlay with mineral oil (Sigma) in a 1.5-mL centrifuge tube
4. Incubate at 65°C in a heating block overnight
5. Incubate at 98°C for 10 min and transfer lysate to a new 1.5mL tube (*see* **Note 2**)
6. Set up 50-μL PCR reactions with 45 μL of PCR Supermix, 2 μL of viral template, 1 μL of forward, and 1 μL of reverse primers (each at a concentration of 100 μM) for the specific viral gene to be amplified, and 1 μL of nuclease-free water (*see* **Note 3**)
7. Set up PCR thermal cycling with 1 cycle of 94°C for 3 min; and 30 cycles of 94°C for 1 min, 58°C for 1 min and 72°C for 2 min (*see* **Note 4**)
8. Load 20% of PCR products on 3% agarose gels for electrophoresis
9. Stain gel with ethidium bromide for 10 min on a rocking platform, destain for 15 min with dH_2O and photograph under an ultraviolet light trans-illuminator
10. Carefully check for successful amplification of each viral gene based on the size of the product on the agarose gel before continuing to the next step (**Fig. 1**) (*see* **Note 5**)

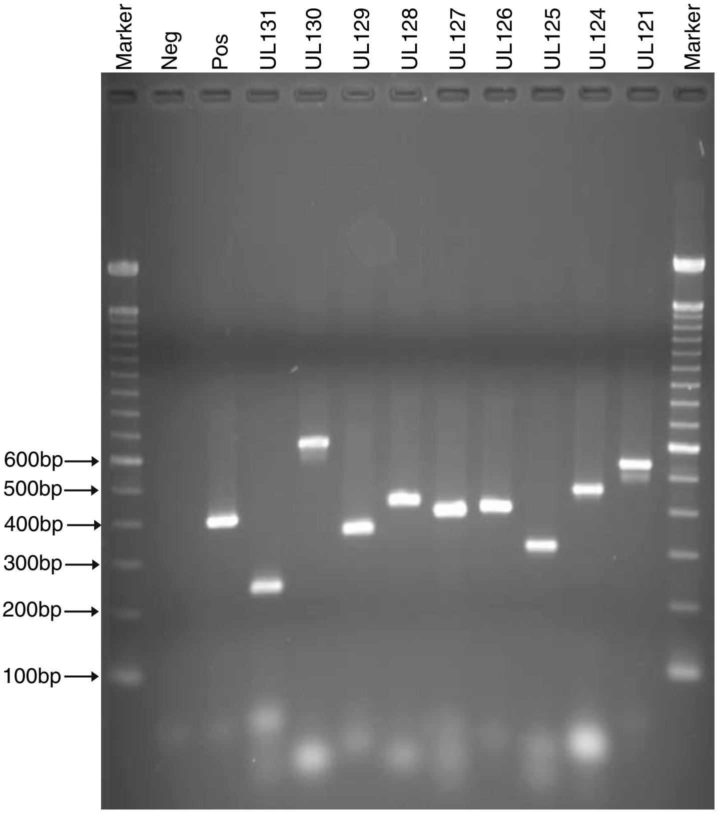


Fig. 1. Confirmation of successful amplification of specific HCMV gene sequences by agarose gel electrophoresis. Ethidium bromide-stained agarose gel showing PCR products from nine HCMV genes in the U_L region of the HCMV genome. A positive control for the presence of amplifiable HCMV DNA (Pos) using well-characterized primers HCMV *ie*1/*ie*2 region gene primers IEP3C/IEP4BII (band size 387 bp) and a negative control (Neg) containing no DNA template were included. 100-bp molecular weight markers were included on both sides of the gel to aid in sizing of bands. Arrows indicate the molecular size markers between 100 and 600 bp. The nine products show strong, discrete bands on the gel, which are ideal for spotting down onto microarray slides.

3.2. Purification and Spotting of Viral Gene Sequences onto Glass Slides

1. Transfer PCR products to Multiscreen PCR_{96} filter plates and place on a Multi-Screen vacuum manifold

2. Apply vacuum suction at 20 inches Hg for 5 min until dried
3. Resuspend the products with 100 μL of milli-Q H_2O to wash the products
4. Repeat vacuum until dried
5. Resuspend products with 40 μL of milli-Q H_2O
6. Transfer products to fresh 96-well round-bottom plates and store at –80°C
7. When ready for printing, thaw PCR products at room temperature, and air dry under vacuum
8. Resuspend in 12 μL of MWG Spotting Buffer (MWG Biotech) and transfer to a 384-well plate (Millipore)
9. Set up program for printing and place in the ESI Microarrayer Robot hood for printing. Follow manufacturer's instructions for printing of microarrays (*see* **Note 6**)
10. Store microarrays in a dust free, dry container at room temp (*see* **Note 7**)

3.2.1. Post-Processing of Viral Gene Microarrays

Prior to hybridization, microarrays need to be processed to remove unbound nucleic acids and printing buffer salts.

1. Use a diamond pen to mark the boundaries of the printed area on the microarray (*see* **Note 8**)
2. Place array inverted over warm 2× SSC for a few seconds to rehydrate the spots and immediately place onto the surface of a clean heating block at 90°C for 5 sec
3. Fix DNA onto glass slides by placing microarray slides into a UV Stratalinker with 600 μJ UV light applied
4. Place slides in a slide rack and wash slides in 1× SSC/0.05% SDS for 30 sec in a glass slide chamber by plunging up and down, and then in 0.06× SSC for 30 sec in another slide chamber
5. Have 6.0 g of succinic anhydride dissolved in 335 mL 1-methyl-2-pyrrolidinone in a slide chamber, and add in 15 mL of sodium borate as soon as the succinic anhydride dissolves (*see* **Note 9**)
6. Immerse microarrays immediately into solution in step 5 and plunge vigorously for 60 sec, and then place on a rocking platform for 10–15 min
7. Place microarrays in boiling water (turn off heat prior to adding microarrays to avoid bubbles) in a 1-liter glass beaker for 2 min, and then transfer into 95% ethanol solution in a slide chamber (*see* **Note 10**)
8. Wash slides in 95% ethanol solution for 15 sec
9. Centrifuge microarrays in a plate centrifuge for 2 min at 75 × g to dry the microarray slides
10. Proceed with hybridization (**section 3.4**) or store dried microarrays in dust-free slide boxes at room temperature (*see* **Note 11**).

3.3. RNA Extraction, Amplification, and Labeling of cDNAs

3.3.1. RNA Extraction

Total RNA is extracted from cells using the RNAqueous kit (Ambion).

1. Lyse mock- and HCMV-infected cell pellets with 700 μL of Lysis/Binding Solution and vortex (*see* **Note 12**)
2. Add an equal volume of 64% ethanol and mix by inverting 5 times
3. Transfer 700 μL of the mix to a Filter Cartridge inserted in an Elution Tube and microcentrifuge at maximum speed (20,800 × g) for 1 min
4. Discard the flow-through
5. Transfer the remaining mixture to the same filter and centrifuge for 1 min at maximum speed
6. Add 500 μL of Washing Solution #1 to the cartridge and centrifuge for maximum speed for 1 min
7. Discard flow through and repeat step 5, then centrifuge an additional minute to allow residues to flow through
8. Transfer Filter Cartridge to a new Elution Tube
9. Apply 50 μL of Elution Buffer heated to 100°C (*see* **Note 13**) to the center of the filter and centrifuge for 30 sec
10. Repeat step 9 for a second elution
11. Add 10 μL of ammonium acetate, and add 1 μL linear acrylamide and mix gently
12. Add 200 μL of ice-cold 100% ethanol and mix well
13. Precipitate RNA overnight by placing into a –80°C freezer
14. Next day, thaw and centrifuge RNA samples in a pre-cooled 4°C microcentrifuge at maximum speed for 20 min
15. Carefully remove supernatant with fine tip pipettes (*see* **Note 14**) and wash pellet with 500 μL of ice-cold 70% ethanol and microcentrifuge for 15 min at maximum speed at 4°C
16. Remove supernatant and allow to air dry for 10–15 min (*see* **Note 15**)
17. Resuspend RNA pellet in nuclease-free water and store at –80°C

3.3.2. Assessment of RNA Quantity and Quality

1. Dilute 1 μL of RNA sample in 49 μL of nuclease-free water and acquire absorbance readings for 260nm and 280nm using a Eppendorf Biophotometer (Eppendorf, Hamburg, Germany) (*see* **Note 16**)
2. Make up a 1% (w/v) MOPS agarose gel (*see* **Note 17**) and electrophorese using 1 μL of RNA sample with loading dye, at 100 V for 30 min
3. Stain gel with ethidium bromide for 10 min and de-stain in dH_2O for 15 min
4. Photograph gel under UV trans-illumination (*see* **Note 18**)

3.3.3. Labeling of RNA by Reverse Transcription by Direct Labeling

If total RNA yields of 25 μg or more are obtained for both mock- and HCMV-infected samples, labeling can proceed as described in this section. However, linear amplification of RNA is required prior to labeling if total RNA yields are low, i.e., 1–5 μg. Linear amplification is described in **section 3.3.4.**

1. Combine 13.4 μL of total RNA (25 μg) with 2 μL of 3 μg/μL oligo-dT primers
2. Heat to 65°C for 10 min, and then quench on ice for 2 min
3. Prepare reverse transcription master mix, containing for each sample: 6 μL of 5× First Strand Buffer, 3 μL of 0.1M DTT, 0.6 μL of 10 mM unlabeled dNTP mix, and 2 μL of 200 U/μL Superscript II
4. Add 3 μL of Cya-3-dUTP to the mock infected RNA sample reaction and 3 μL of Cya-5-dUTP to the virus infected RNA sample reaction
5. Incubate at room temperature wrapped in aluminium foil for 2 h (*see* **Note 19**)
6. Add 15 μL of 0.1N NaOH and incubate at 70°C for 10 min to degrade any remaining RNA
7. Add 7.5 μL of 0.1N HCl to neutralize the reaction mixture
8. Add 400 μL 1× TE to each sample and transfer to Microcon YM-30 filter
9. Microcentrifuge samples at maximum speed for 11–15 min at room temperature until 10–20-μL volume remains on the filter (*see* **Note 20**)
10. Discard flow-through and add 450 μL 1× TE buffer to the filter and repeat step 9
11. Invert filter into a new collection tube and microcentrifuge at maximum speed for 1 min to collect the labeled probes
12. Combine the 2 samples (mock and infected) in a single Microcon YM-30 filter, and add 450 μL 1× TE buffer to the labeled mixtures together with 20 μL of Cot-1 Human DNA, 2 μL of poly-A RNA, and 2 μL of tRNA (*see* **Note 21**)
13. Microcentrifuge for 11–15 min at maximum speed to bring the volume of the labeled mixture to less than 20 μL
14. Recover labeled mixture by repeating step 11
15. The labeled mixture is now ready for hybridization to a microarray (*see* **section 3.4**)

3.3.4. Linear Amplification of RNA by in Vitro Transcription

If total RNA yields are low, i.e., 1–5 μg, linear RNA amplification needs to be performed before labeling and hybridization. This is carried out using the MessageAmp II aRNA amplification kit from Ambion:

3.3.4.1. Double-Stranded cDNA Synthesis

1. Mix 1–5 μg of total RNA from mock- and HCMV-infected cells in 10 μL nuclease-free water with 1 μL of T7 Oligo(dT) primer and incubate for 10 min at 70°C in a thermal cycler

2. Add 8 µL of reverse transcription master mix (2 µL 10× First Strand Buffer, 4 µL dNTP Mix, 1 µL RNase inhibitor, 1 µL ArrayScript) to each reaction
3. Incubate at 42°C in a thermal cycler for 2 h
4. Add 80 µL of Second Strand Master Mix to each sample (63 µL nuclease-free water, 10 µL 10× Second Strand Buffer, 4 µL dNTP Mix, 2 µL DNA Polymerase, 1 µL RNase H)
5. Incubate at 16°C in a thermal cycler for 2 h

3.3.4.2. Double-Stranded cDNA Purification

6. Transfer reaction mixture into 1.5-mL eppendorf tubes and add 250 µL cDNA Binding Buffer to each sample and mix well
7. Place sample mix into a cDNA Filter Cartridge and microcentrifuge for 1 min at 10,000 × g
8. After flow-through is discarded, add 500 µL of Wash Buffer and microcentrifuge for 1 min at 10,000 × g
9. Discard flow-through and microcentrifuge for an additional minute at 10,000 × g to remove traces of Wash Buffer
10. Transfer filter into fresh cDNA Collection tubes
11. Elute cDNA with 10 µL nuclease-free water at 50°C applied to the center of the filter, and microcentrifuge at 10,000 × g for 1 min (*see* **Note 22**)
12. Perform a second elution by repeating step 11

3.3.4.3. Generation of Amplified RNA (aRNA) by in Vitro Transcription

The in vitro transcription step results in linear amplification, generating amplified RNA (aRNA).

13. To the purified cDNA samples, add 24 µL of IVT Master Mix (4 µL each of 75 mM T7 ATP, CTP, GTP, TTP Solutions, 4 µL of T7 10× Reaction Buffer, 4 µL of T7 Enzyme Mix)
14. Incubate overnight at 37°C for 8–12 h (*see* **Note 23**)
15. Add 60 µL of nuclease-free water after the 8–12 h incubation
16. Mix aRNA sample with 350 µL of aRNA Binding Buffer, and then mix with 250 µL of 100% ethanol
17. Transfer mixture to aRNA Filter Cartridge and microcentrifuge at 10,000 × g for 1 min
18. Discard flow-through and add 650 µL Wash Buffer to the Filter Cartridge, and microcentrifuge at 10,000 × g for 1 min
19. Discard the flow-through, and microcentrifuge an additional minute at 10,000 × g
20. Transfer Filter Cartridge into a fresh aRNA Collection Tube, apply 100 µL of nuclease-free water preheated at 50°C
21. Leave at room temperature for 2 min, and microcentrifuge for 2 min at 10,000 × g
22. Precipitate the aRNA by adding 10 µL of 5M ammonium acetate, 275 µL of 100% ethanol, mix well and incubate overnight at –80°C

23. Next day, centrifuge RNA samples in a pre-cooled 4°C microcentrifuge at maximum speed for 20 min
24. Carefully remove supernatant with fine tip pipettes and wash pellet with 500 μL of ice-cold 70% ethanol and microcentrifuge for 15 min at maximum speed at 4°C
25. Remove supernatant and allow to air dry for 10–15 min (*see* **Note 15**)
26. Resuspend RNA pellet in nuclease-free water and store at –80°C
27. Assess aRNA quality by 1% agarose gel electrophoresis (*see* **Note 24**) **(Fig. 2)**.

3.3.5. Labeling of Amplified RNA

1. Prepare 1–3 μg of aRNA in 13.4 μL nuclease-free water
2. Mix with 2 μL of random primers, heat to 65°C for 10 min and quench on ice for 2 min
3. Follow step 3 onward in **section 3.3.3** to complete the labeling procedure
4. Samples are now ready for hybridization to a microarray

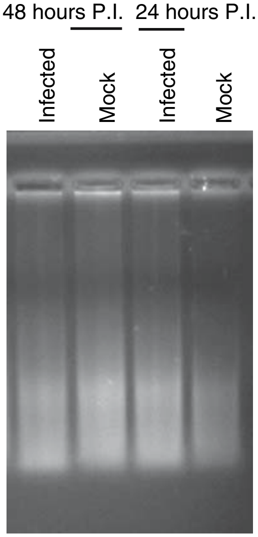


Fig. 2. Generation of aRNA by linear amplification. Ethidium bromide-stained 1% agarose gel showing separation under denaturing conditions of aRNA samples from mock infected and HCMV strain Towne infected myeloid progenitor cells at 24 and 48 h post infection (P.I.). Successful generation of aRNA is indicated by the presence of a characteristic nucleic acid smear representing a large range of amplified poly-A-containing mRNAs, with the majority of aRNAs ranging in size from 200 bp to 1,800 bp.

3.4. Microarray Hybridization and Washing

1. Combine labeled cDNAs mixture from mock- and HCMV-infected samples generated from either total RNA (**section 3.3.3**) or amplified RNA (**section 3.3.4**) and adjust to 20 μL with 1× TE buffer
2. Add 4.25 μL of 20× SSC and 0.75 μL of 10% SDS to the labeled mixture
3. Denature by heating for 2 min at 100°C
4. Incubate at room temperature for 15 min covered in aluminium foil
5. Carefully pipette labeled cDNAs mixture to the centre of a HCMV gene microarray printed area and cover it with a 22 × 50-mm coverslip (*see* **Note 25**)
6. Place microarray in a Glass Array Hybridization Cassette, and carefully place a 6 μL spot of nuclease-free water each in the two insets inside the cassette. Seal the cassette (*see* **Note 26**)
7. Place the cassette in a water bath overnight at 65°C (*see* **Note 27**)
8. After hybridization, remove microarrays from the cassette
9. Gently wash in a slide chamber filled with 2× SSC/0.1% SDS until the cover slip falls off (use fine forceps to assist removing the cover slip), and agitate for a further 15 sec
10. Transfer to a fresh staining dish and wash by gentle agitation in 1× SSC for 30 sec, and then transfer to 0.2× SSC and wash for another 30 sec
11. Dry microarrays by centrifugation in a plate centrifuge for 5 min at 75 × g
12. Place in a dust-free slide box in the dark until scanning (*see* **Note 28**)

3.4.1. Scanning and Data Analysis

1. Place microarray face down into a GenePix^{TM} 4000B scanner (Axon Instruments Inc.)
2. Using the GenePix Pro software, adjust laser power for 635nm and 532nm lasers to a PMT gain setting of 600 (*see* **Note 29**)
3. Perform a low-resolution preview scan at 40 μm
4. Using the "scan area" function, draw a box around the printed area where the gene quadrants are located (*see* **Note 30**)
5. Perform a high-resolution scan at 5 μm
6. Zoom in so that only the 94 assorted human genes are visible on the screen
7. Observe overall 635-nm and 532-nm intensities in the histogram tab of the main window
8. Adjust laser voltages so that the overall intensity for the assorted 94 human genes quadrant for the 2 lasers to match each other (*see* **Note 31**)
9. Overlay the template grid (.GAL file) onto the scanned microarray image and align over the gene spots (*see* **Note 32**)
10. The align feature will find and flag gene spots automatically. Flagged spots are defined as Bad (–100), Absent (–75), and Not Found (–50) or Unflagged (0) by the GenePix Pro software (*see* **Note 33**)
11. Once spots are aligned, utilize the integrated GenePix Pro software "Analyse tool" to extract the numerical values for each gene spot. The median foreground pixel

intensity and morphological opening background estimation are the recommended analysis options.

12. Save the image file (.TIFF and .JPEG), the settings (layout) file (.GPS), and the results file (.GPR)
13. Import the results files (.GPR) into the program R (version 1.9.1 was used for the following steps) ***(26)*** and install the additional bioconductor packages. An in-house Graphical User Interface to simplify access to the tools within Bioconductor and automate the analysis procedure was used for the processing of the data. A simplified outline of the basic tools and steps are described below (*see* **Note 34**).
14. Input .GPR data files into R workspace using the marray package together with the details of the layout of microarrays used, including the foreground (F635, F532) and background (B635, B532) intensities, the number of gene spots per row (12) and column (16), and the number of quadrants (2 × 2)
15. Define gene type on the microarray: viral genes are labeled VIRAL, human genes as HUMAN, and salt spots as BUFFER (*see* **Note 35**)
16. After the data has been inputted into R, use the threshold filtering function to convert any intensity values less than 0 to a value of 1.0, although the morphological opening function now available in GenePix Pro means that negative intensities are no longer an issue and so do not need to be converted. Confirm the quality of each array using diagnostic plots and remove low quality arrays (*see* **Note 36**)
17. Filter out gene spots that have a flag value less than 0 (*see* **Note 37**)
18. Normalize the data with the marray package for all replicates within one timepoint using the loess normalization method ***(27)*** using the subset "HUMAN," i.e., the set of 94 assorted human genes
21. To calculate the mean (*M*), standard deviation (*s*), and the 90th percentile (*a*) values of the normalized data (in log2 values) for each viral gene across the replicates, export the normalized data using the export function command in the program R into Microsoft Excel. Alternatively, these calculations can be calculated and ranked in R.
22. Using the limma package calculate the penalized t-statistic value for each gene across the number (*n*) of replicates using the formula described by Efron and colleagues ***(28)*** and Smyth and colleagues ***(29)*** (*see* **Note 38**):

$$t = \frac{M}{(a+s)/\sqrt{n}}$$

23. Rank viral genes that has a penalized t-statistic, i.e., a log2 value of greater than or equal to 1.0 is considered to be expressed (*see* **Note 39**).

3.5. Validation of Microarray Results using RT-PCR

1. Using 1 μg of total RNA in 7 μL of nuclease-free water, add 1 μL of RQ1 Buffer, and 1 μL of RQ1 DNase enzyme
2. Mix well and incubate at 37°C for 30 min

3. Stop the DNase treatment by adding 1 μL of RQ1 Stop buffer and incubate at 65°C for 10 min (*see* **Note 40**)
4. Set up 20-μL reverse transcription reactions in 1.5-mL nuclease-free tubes by adding the following:

   a. 4 μL of 5× First Strand Buffer
   b. 2 μL of 0.1M DTT
   c. 1 μL RNaseOUT (40U/μL)
   d. 0.4 μL of 25 mM dNTPs
   e. 0.3 μL of 3 μg/μL Random Primers
   f. 8.8 μL of nuclease-free water

5. Add 2.5 μL DNase-treated RNA to each reaction
6. For each RNA sample, include 2 reactions: +RT or –RT (*see* **Note 41**)

   a. +RT: add 1 μL of Superscript II reverse transcriptase
   b. –RT: add 1 μL of nuclease-free water

7. For each RT setup, include a positive (using cDNA from productively infected HFFs with HCMV as template) and negative control (nuclease-free water as template)
8. Incubate reactions at room temperature for 15 min, then on a heating block for 42°C for 1 h
9. Stop the RT reactions by incubating at 70°C for 15 min (*see* **Note 42**)
10. Set up 50 μL PCR reactions for each RT reaction as follows: 5 μL of 10× PCR Reaction Buffer

    a. 2 μL of 50 mM $MgCl_2$
    b. 0.4 μL of 25 mM dNTPs
    c. 1 μL each of gene-specific forward and reverse primers
    d. 37.1 μL of nuclease-free water
    e. 2.5 μL of RT reaction
    f. 1 μL of Platinum Taq (5U/μL)

11. Perform PCR thermal cycling using the conditions: 94°C for 3 min, 40 cycles × [94°C for 1 min, Y°C for 1 min, 72°C for 1 min], followed by 72°C for 10 min (where Y is the annealing temperature for the gene primer pair in interest)
12. Load 20% of PCR products with loading dye onto 2% agarose gels and electrophorese at 100V for 30 min
13. Stain the gel with ethidium bromide for 10 min, destain for 15 min, and photograph the gel under UV-transillumination
14. Check for the correct sized band on the gel

Data to confirm the presence or absence of HCMV gene expression in myeloid progenitors by RT-PCR are shown in **Table 1**. In this experiment, RNA was extracted from latently infected myeloid progenitors on days 1, 2, 3, 5, and 11 after infection was examined by RT-PCR for the expression of

Table 1
Confirmation of the presence or absence of HCMV gene expression in myeloid progenitors by RT-PCR

	Day after infection									
HCMV GENE	1		2		3		5		11	
	RT–	RT+	RT–	RT+	RT–	RT+	RT–	RT+	RT–	RT+
UL68	–	+	–	+	–	+	–	+	–	+
UL50	–	–	–	–	–	–	–	–	–	–
UL120	–	–	–	–	–	–	–	–	–	–

HCMV UL68, UL50, and UL120 transcripts for 40 cycles of amplification. Consistent with the microarray-based analysis of viral gene expression, UL68 transcripts, but not UL50 or UL120 transcripts were detected in the presence of reverse transcriptase (RT+). Omission of reverse transcriptase (RT–) did not yield amplified products, confirming that amplification was derived from an RNA template rather than any contaminating viral DNA.

3.6. Results and Conclusions

This report has described the construction and application of HCMV gene-specific microarrays to interrogate the HCMV transcriptome during experimental productive infection of HFFs and also during the establishment phase of latent infection in myeloid progenitor cells. In **Fig. 3** microarray-based detection of HCMV gene transcription during productive infection of human foreskin fibroblast cells is shown. RNAs from either mock infected or HCMV-infected fibroblasts were labeled with either Cya-3 (mock) of Cya-5 (infected) by reverse transcription and hybridized HCMV gene microarrays. After washing, microarrays were scanned with a dual laser scanner to simultaneously excite both Cya-3 (532 nm) and Cya-5 (635nm) and an overlaid image was generated. Top panel: As early as 2 hours post infection, viral gene expression was detected from the *UL122/UL123* gene region (box), which encodes the first genes expressed by HCMV upon initiation of a productive infection. Bottom panel: At 3 days post-infection, expression from a majority of genes was detected. The positions of HCMV gene spots and human gene control spots are indicated.

Fig. 4 provides a summary of the microarray-based detection of HCMV gene transcription during the establishment of a non-productive, latent infection in primary human myeloid progenitor cells. CD34$^+$/CD33$^+$ myeloid progenitors

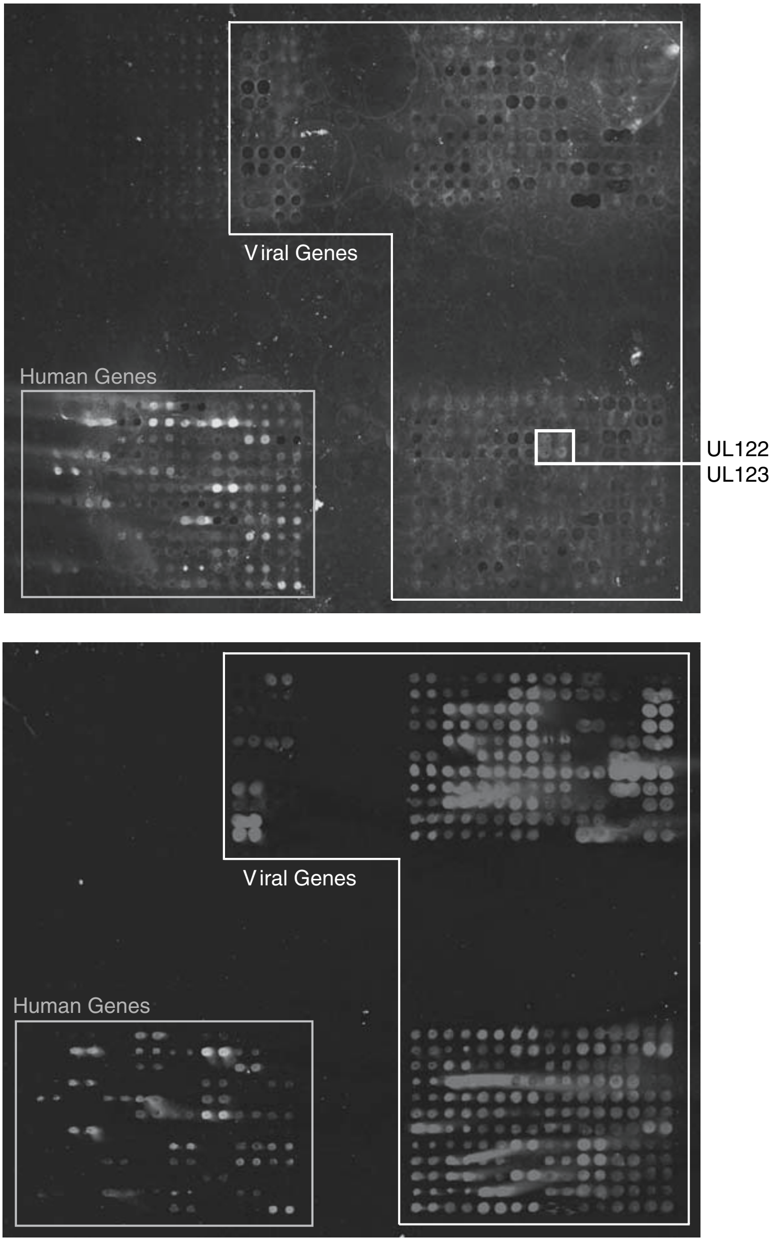


Fig. 3. Microarray-based detection of HCMV gene transcription during productive infection of human foreskin fibroblast cells (HFFs).

	Time post infection				
HCMV gene	**Day 1**	**Day 2**	**Day 3**	**Day 5**	**Day 11**
RL2	■	■	■	■	
RL3	■	■	■	■	
RL4	■	■	■	■	
RL5	■				
RL6	■	■	■		
RL7	■	■	■	■	■
RL11			■		
UL5	■				
UL23	■				
UL30	■			■	
UL39	■				
UL53			■		
UL56			■		
UL62	■				
UL64	■				
UL66			■		
UL67	■		■		■
UL68	■	■	■	■	■
UL71		■			
UL81	■				
UL89	■				
UL90		■			
UL93		■			
UL99	■		■		
UL108	■	■	■		■
UL110	■		■	■	■
UL111	■		■	■	
UL111A					■
UL115					■
UL123			■		
UL132	■				
UL147	■				
UL153	■	■	■	■	
UL154	■	■	■	■	
US12	■				
US28			■		
US32				■	

Fig. 4. Summary of the microarray-based detection of HCMV gene transcription during the establishment of a non-productive, latent infection in primary human myeloid progenitor cells.

were infected with HCMV and RNAs extracted over an 11 day time course to identify HCMV gene expression as the virus established a latent infection within these cells. Due to low cell numbers and RNA yields, RNAs were amplified by linear amplification prior to labeling and hybridization to HCMV gene microarrays. Expression of a subset of 37 HCMV genes represented on the microarrays was detected and these are listed, with shaded boxes indicating the time point(s) post-infection at which expression was detected.

4. Notes

1. HFFs cultured in Dulbecco's modified Eagle's medium supplemented with 10% FCS (DMEM) at 80–90% confluency in a T75 cm^2 culture flask (Beckon Dickinson, San Jose, CA) were washed with phosphate buffered saline, infected with HCMV at a multiplicity of infection (MOI) of 1 and infection allowed to proceed for 5 days.
2. This treated lysate, containing viral DNA can now be used as the template for PCR amplification of individual HCMV genes.
3. Positive control used primers - IEP3C and IEP4BII ***(25)***, while negative control contains the same primers with 2 μL nuclease-free water added instead of DNA template.
4. If these cycle parameters do not produce a successful amplification of any particular gene, optimization of cycling conditions is required, including varying the number of cycles (30-40 cycles) and the annealing temperature (55–65°C).
5. Strong amplification of a single gene product at the correct size is required for a gene product to be acceptable for printing onto the microarray. Multiple bands or bands at the incorrect size usually indicate non-specific amplification.
6. Briefly, the 384-well plate containing all amplified HCMV genes together with 94 assorted human genes was placed in a dust-free hood containing the microarrayer robot. The printing program was setup such that duplicate spots are deposited adjacent to each other.
7. Dried microarrays can be stored for approximately 12 months in a dust-free container, in a dark, dry environment.
8. It is important to mark the boundaries of the spots on printed microarray because although at this stage the spots will be visible due to the presence of salts, they will not be visible after post-processing.
9. The succinic anhydride solution needs to be made fresh and ready by the end of step 4: the 1× SSC/0.05% SDS wash.
10. Have the boiling water bath and 95% ethanol prepared before step 7, so that the microarrays can be immediately washed in these solutions.
11. After microarrays are dried from centrifugation, immediately place them into a dust-free slide box. Care needs to be taken to avoid any dust particles landing on the microarrays, which may lead to increase background after hybridization and scanning.

12. Vortex until clear, if there is still cell debris or cloudiness, add more Lysis/Binding solution until clear.
13. Heat up Elution Buffer before starting RNA extraction.
14. Use a 1,000-μL pipette to remove as much supernatant as possible, then switch to a 200 μL pipette and remove more of the residue supernatant, followed by a quick spin, and then removal of the remaining liquid using a 20 μL pipette.
15. Leave Elution (or Collection) tube lid(s) open and create a "tent" with aluminium foil to allow air drying whilst minimizing dust particles from entering the tubes.
16. Use the same nuclease-free water used to resuspend RNA as blank for spectrophotometric analysis. An A260:A280 ratio of 1.8–2.0 indicates good quality RNA.
17. Cast a 1% agarose gel in the presence of 1× MOPS (3-[N-Morpholino] propanesulfonic acid, Sigma).
18. Well-defined 28S and 18S ribosomal RNAs present on the gel in a 2:1 ratio are indicative of high-quality RNA.
19. Adding an additional 1 μL of Superscript II enzyme after 1 h of incubation may sometimes improve labeling/cDNA synthesis.
20. Approximately 10–20 μL would appear as a "crescent moon" shape of labeled mixture along one side on top of the filter .
21. Cot-1 DNA, poly-A RNA and tRNA are used to block non-specific binding of probes.
22. Pre-heat nuclease-free water before starting the purification procedures.
23. According to manufacturer's notes, optimal amplification occurs after 10–12 h of in vitro transcription.
24. Successful generation of aRNA is indicated by the presence of a characteristic nucleic acid smear representing a large range of amplified poly-A-containing mRNAs (**Fig. 2**).
25. Hold coverslip on ends with gloved fingertips, applying very slight pressure towards the middle to bend the cover slip, and place onto labeled mixture on the microarray. If there are air bubbles, gently use a pipette tip to push them out by applying pressure to the upper side of the cover slip.
26. Water added to the insets is used to maintain humidity during hybridization. Tighten each screw on the cassette evenly.
27. Pre-heat water bath to 65°C before labeling reactions are completed. If any leaking occurs in the cassette after immersed in the water bath (indicated by release of air bubbles), be prepared to immediately remove the cassette and fix leak.
28. For optimal results, scan immediately, although slides can be left in the slide box, in the dark for up to a week before scanning. Signals will deteriorate over time.
29. A PMT gain setting of 600 is the default for the Axon scanner and generally accepted setting for both lasers for an initial scan.
30. The boxed area defines where the scan will be performed. The box function may differ between different types of scanners.
31. The assumption here is that on average, the human gene will not be significantly altered in their expression between mock and HCMV-infected samples. Using the histogram option in the associated GenePix Pro software, increase or decrease

the 2 laser PMT settings so that the red and green lines are aligned (representing the range, 635nm and 532nm laser intensities, respectively). Also, ensure that the whole x-axis scale is used, i.e. both lines extend almost all the way to the right end of the x-axis to make use of the full dynamic range of the scanner. Repeated preview scans maybe required to match the two laser settings.

32. Once roughly aligned, use the "auto-align, auto find" features option to allow the program to align accurately each spot.
33. An "absent" spot is defined in the .GAL file when no gene was printed. A spot is assigned a "not found" flag when the alignment fails to find the spot. The user can manually re-center the spot and re-align the feature. "Good" and "bad" flags are set manually by the user. Normal gene spots are unflagged and have a value of 0. The user will need to screen the microarray to ensure all the spots are distinctly segmented with the grid for each spot.
34. R software is freely available at www.r-project.org, as documented in Gentleman et. al. ***(26)***. Bioconductor packages to expand the functionality of R can be downloaded at freely available at www.bioconductor.org (Fred Hutchinson Cancer Research Centre, Seattle, WA). A generally available graphical user interface to simplify access to some of the Bioconductor tools used in the analysis is available through the limmaGUI package.
35. This allows the division of the gene spots into three different groups that will aid further analysis. The gene types can be defined manually by modifying the .GAL file to include an additional column called "Control" and entering VIRAL, HUMAN or BUFFER for each gene. An automated function in the marray package can also be used to define the control types.
36. Arrays may be of low quality as a result of a failed hybridization, low quality RNA, high background or scanning problems. Diagnostic plots are available through the marray package.
37. A flag less than 0 indicate genes that are flagged "bad," "absent," or "not found," which may contain high background or irrational values that could influence the outcome of the analysis and are therefore replaced with an NA
38. Smyth and colleagues ***(29)*** argued that the ordinary t-statistic for microarray data analysis is not ideal because a large t-statistic can be the result of an unrealistically small standard deviation. Thus, genes that have a small variance but not a high mean value could be defined as being expressed. Therefore, a compromise between the mean value and the standard deviation is needed. Efron's study ***(28)*** illustrated that using a non-parametric Empirical Bayesian approach to microarray data analysis for the t-statistic, using an "*a*" value as the 90th percentile represents higher values, which there will be less information loss from the full data to the summarized statistic value. This approach leads to an estimated "log-odds" that each gene is defined as being differentially expressed.
39. This converts to a ratio of intensities between 635- and 532-nm channels as 2.0, and has been used in our publication ***(30)***. This falls into the general criteria of 2.0–3.0 defining a viral gene to be expressed from previous studies ***(18,20,23,31)***.
40. This now contains DNase-treated RNA.

41. Adding or omitting reverse transcriptase can determine if any DNA contamination is present in the RNA sample.
42. The reactions now contain cDNA.

Acknowledgments

The authors would like to thank Chris Bye and Karen Bythe for their assistance with microarray data analyses.

References

1. Mocarski, E. S., and Courcelle, C. T. (2001) Cytomegaloviruses and their replication, in *Field Virology* (Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B., et al., eds.), 4th ed. Vol 2, Lippincott Williams and Wilkins, Philadelphia, PA, pp. 2629–2673.
2. Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., et al. (1990). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr. Top. Microbiol. Immunol.* **154**, 125–69.
3. Cha, T. A., Tom, E., Kemble, G. W., Duke, G. M., Mocarski, E. S., and Spaete, R. R. (1996) Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J. Virol.* **70**, 78–83.
4. Murphy, E., Rigoutsos, I., Shibuya, T., and Shenk, T. E. (2003) Reevaluation of human cytomegalovirus coding potential. *Proc. Natl. Acad. Sci. U S A***100** , 13585–13590.
5. Dolan, A., Cunningham, C., Hector, R. D., Hassan-Walker, A. F., Lee, L., Addison, C., et al. (2004) Genetic content of wild-type human cytomegalovirus. *J. Gen. Virol.***85** , 1301–1312.
6. Murphy, E., Yu, D., Grimwood, J., Schmutz, J., Dickson, M., Jarvis, M. A., et al. (2003) Coding potential of laboratory and clinical strains of human cytomegalovirus. *Proc. Natl. Acad. Sci. U S A* **100**, 14976–14981.
7. Harari, A., Zimmerli, S. C., and Pantaleo, G. (2004) Cytomegalovirus (CMV)-specific cellular immune responses. *Hum. Immunol.* **65**, 500–506.
8. Sinclair, E., Black, D., Epling, C. L., Carvidi, A., Josefowicz, S. Z., Bredt, B. M., et al. (2004) CMV antigen-specific CD4+ and CD8+ T cell IFN gamma expression and proliferation responses in healthy CMV-seropositive individuals. *Viral Immunol.* **17**, 445–454.
9. Stagno, S., Pass, R. F., Dworsky, M. E., Henderson, R. E., Moore, E. G., Walton, P. D., et al. (1982) Congenital cytomegalovirus infection: the relative importance of primary and recurrent maternal infection. *N. Engl. J. Med.* **306**, 945–949.
10. Stagno, S., Pass, R. F., Cloud, G., Britt, W. J., Henderson, R. E., Walton, P. D., et al. (1986) Primary cytomegalovirus infection in pregnancy. Incidence, transmission to fetus, and clinical outcome. *JAMA* **256**, 1904–1908.

11. Demmler, G. J. (1991) Infectious Diseases Society of America and Centers for Disease Control. Summary of a workshop on surveillance for congenital cytomegalovirus disease. *Rev. Infect. Dis.* **13**, 315–329.
12. Pass, R. F. (2001) Cytomegalovirus, in *Fields Virology* (Knipe, D. M., Howley, P. M., Griffith, D. E., Lamb, R. A., Martin, M. A., Roizman, B., et al., ed.), Lippinocott Williams and Wilkins, Philadelphia, PA, pp. 2675–2705.
13. Stanier, P., Taylor, D. L., Kitchen, A. D., Wales, N., Tryhorn, Y., and Tyms, A. S. (1989) Persistence of cytomegalovirus in mononuclear cells in peripheral blood from blood donors. *Br. Med. J.* **299**, 897–898.
14. Taylor-Wiedeman, J., Sissons, J. G., Borysiewicz, L. K., and Sinclair, J. H. (1991) Monocytes are a major site of persistence of human cytomegalovirus in peripheral blood mononuclear cells. *J. Gen. Virol.* **72**, 2059–2064.
15. Mendelson, M., Monard, S., Sissons, P., and Sinclair, J. (1996) Detection of endogenous human cytomegalovirus in CD34+ bone marrow progenitors. *J. Gen. Virol.* **77**, 3099–3102.
16. Soderberg-Naucler, C., Fish, K. N., and Nelson, J. A. (1997) Reactivation of latent human cytomegalovirus by allogeneic stimulation of blood cells from healthy donors. *Cell* **91**, 119–126.
17. Reeves, M. B., MacAry, P. A., Lehner, P. J., Sissons, J. G., and Sinclair, J.H. (2005) Latency, chromatin remodeling, and reactivation of human cytomegalovirus in the dendritic cells of healthy carriers. *Proc. Natl. Acad. Sci. U S A* **102**, 4140–4145.
18. Chambers, J., Angulo, A., Amaratunga, D., Guo, H., Jiang, Y., Wan, J. S., et al. (1999) DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J. Virol.* **73**, 5757–5766.
19. Stingley, S. W., Ramirez, J. J., Aguilar, S. A., Simmen, K., Sandri-Goldin, R. M., Ghazal, P., et al. (2000) Global analysis of herpes simplex virus type 1 transcription using an oligonucleotide-based DNA microarray. *J. Virol.* **74**, 9916–9927.
20. Ebrahimi, B., Dutia, B. M., Roberts, K. L., Garcia-Ramirez, J. J., Dickinson, P., Stewart, J. P., et al. (2003) Transcriptome profile of murine gammaherpesvirus-68 lytic infection. *J. Gen. Virol.* **84**, 99–109.
21. Jones, J. O., and Arvin, A. M. (2005) Viral and cellular gene transcription in fibroblasts infected with small plaque mutants of *varicella-zoster virus. Antiviral Res.***68** , 56–65.
22. Kennedy, P. G., Grinfeld, E., Craigon, M., Vierlinger, K., Roy, D., Forster, T., et al. (2005) Transcriptomal analysis of varicella-zoster virus infection using long oligonucleotide-based microarrays. *J. Gen. Virol.***86** , 2673–2684.
23. Lua, D. T., Yasuike, M., Hirono, I., and Aoki, T. (2005) Transcription program of red sea bream iridovirus as revealed by DNA microarrays. *J. Virol.***79** , 15151–15164.
24. Aguilar, J. S., Devi-Rao, G. V., Rice, M. K., Sunabe, J., Ghazal, P., and Wagner, E. K. (2006) Quantitative comparison of the HSV-1 and HSV-2 transcriptomes using DNA microarray analysis. *Virology* **348**, 233–241.

25. Kondo, K., Kaneshima, H., and Mocarski, E. S. (1994) Human cytomegalovirus latent infection of granulocyte-macrophage progenitors. *Proc. Natl. Acad. Sci. U S A* **91**, 11879–11883.
26. Gentleman, R. C., Carey, V. J., Bates, D.M., Bolstad, B., Dettling, M., Duoit, S., et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.
27. Smyth, G., and Speed, T. (2003) Normalization of cDNA microarray data. *Methods (Duluth)* **31**, 265–273.
28. Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160.
29. Smyth, G. K., Yang, Y. H., and Speed, T. (2003) Statistical issues in cDNA microarray data and analysis. *Methods Mol. Biol.* **224**, 111–136.
30. Cheung, A. K., Abendroth, A., Cunningham, A. L., and Slobedman, B. (2006) Viral gene expression during the establishment of human cytomegalovirus latent infection in myeloid progenitor cells. *Blood* **August 24**, 2006.
31. Ahn, J. W., Powell, K. L., Kellam, P., and Alber, D. G. (2002) Gammaherpesvirus lytic gene expression as characterized by DNA array. *J. Virol.* **76**, 6244–6256.

10

Computer-Assisted Reading of DNA Sequences

Huong Le, Marcus Hinchcliffe, Bing Yu, and Ronald J. A. Trent

Summary

DNA sequencing is increasingly used in a range of medical activities involving DNA diagnostics and research. This is the result of improving technology and cheaper costs. Paradoxically, a greater demand for DNA sequencing has placed additional work on the laboratory because sequencing profiles must be checked visually despite the availability of informatics-based tools in interpreting DNA sequence traces. In this environment it is essential to have more sophisticated software that will allow the sites of known and unknown DNA variants to be quickly identified, as well as providing an objective assessment of quality for the DNA sequence generated. This chapter describes the Applied Biosystems SeqScape® software program (version 2.5) and how it has assisted in the interpretation of DNA sequencing in a DNA diagnostic laboratory.

Key Words: DNA sequencing, computer-assisted reading, mutation.

Abbreviations: kb – kilobase; QV – quality value; RDG – reference data group; ROI – regions of interest; IUB – international union of biochemistry

1. Introduction

One of the successful goals of the Human Genome Project was to develop new platforms and technologies to make DNA sequencing faster and cheaper *(1)*. This has now produced rapidly expanding datasets containing DNA sequence information that have accelerated disease-causing mutation detection, as well as SNP discovery.

There are many DNA-based strategies to detect DNA variants, but the gold standard remains DNA sequencing *(2)*. As such it provides complete

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

information about the location and nature of all DNA variants *(3)*. Improvements in DNA sequencing technologies have meant longer read lengths. Genes can now be sequenced routinely with more accurate base calls and higher quality traces. Reads involving 800 bases of DNA sequence can be routinely obtained using capillary electrophoresis. This has led to a significant improvement in SNP discovery *(4)*. However, high-throughput DNA analytic techniques generate a lot of raw data that need to be interrogated. Many computer software technologies for assembling and analyzing raw data have been developed but despite this, manual visual inspection of DNA sequence traces is still required, particularly if these are to be used for diagnostic or clinical purposes. Furthermore, to ensure reliable results particularly when heterozygous mutations or DNA variants are being sought, both strands of the DNA sequence must be analyzed *(5)*.

Manual comparisons of unknown samples with positive control and wild-type sequences are increasingly taking more time in the diagnostic or research laboratories as the length and number of DNA sequences increase. There are few reliable informatics-based tools for variant identification, result viewing, data tracking, and result reporting. Therefore, to improve the efficiency of DNA sequencing and analysis, more effort is needed to develop better software. As an example of what is needed, we describe the Applied Biosystems SeqScape® software (version 2.5) for automated rapid mutation identification and result reporting.

SeqScape software is a computer-assisted sequencing analysis tool that supports the new KB™ Basecaller (version 1.2) that supports sample files generated from Applied Biosystems instruments. It is a Windows-based program that integrates base calling, sequence assembly, alignment and comparison tools. It can accurately call bases, and assign a quality value per base call thereby allowing the poorer quality segments to be trimmed from the analysis. This enhances the identification process of any mismatched or mixed base, and provides an efficient approach to detecting DNA variants. Hence, SeqScape reduces the time taken for sequence analysis as well as facilitating interpretation. Once mastered, limitations in using SeqScape are few and with each new version different capabilities are added (*see* **Note 1**).

2. Features of Seqscape Software

SeqScape software is designed to provide an improved level of accuracy with the identification of DNA and amino acid variants as well as identifying genotypes, alleles or haplotypes from an available library.

A popular sequencing technique utilizes dye terminator chemistry that allows four reactions to be performed in a single tube and links this to a laser-activated fluorescence detection system. DNA fragments are electrophoresed in capillaries and data are collected as chromatogram or electropherogram files that function as input files for SeqScape software analysis.

SeqScape software contains a number of features that are useful for sequence analysis. These include (1) base-calling, (2) quality value assessment, (3) quality trimming, (4) filtering out poor sequence, (5) sequence assembly, (6) sequence alignment, (7) library match search if there is a link with a known library, (8) identification of pure and mixed bases, and (9) generation of different formats for reporting the results.

Base calling is an essential part of sequencing analysis. It involves an algorithm used to translate the fluorescence signal spectra into the four nucleotide bases associated with the DNA sequence ***(4)***. The current version of SeqScape software uses KB Basecaller, an improved algorithm responsible for more accurate base calling and increasing the read length. KB Basecaller is designed to produce quality values (QV) for every base it calls. The QV is used for the identification of pure and mixed bases contrasting with the original Applied Biosystems base calling algorithm, which only identified pure bases ***(6)***. The KB Basecaller uses equations that are standardized by the Phred base calling software to provide a quality prediction for each base. The accuracy of base calls are calculated from the following parameters: peak spacing, uncalled/called peak ratio and peak resolution ***(7)*** and so provide a quality assessment value per base.

Measuring the error probability of a base call is essential for assuring sequence quality. QV is an estimation of the certainty for a base call ***(6)*** and is derived from the equation QV= $-10 \log_{10}(\text{Pe})$, where Pe is the error probability of base call ***(7)***. A QV of 20 means a predictive sequencing error rate per base of 1.00% (**Table 1**).

Table 1
Illustrating the link between the QV and Pe*

QV	P_e %	QV	P_e %	QV	P_e %
1	79	21	0.79	41	0.008
5	32	25	0.32	45	0.003
10	10	30	0.10	50	0.001
15	3.2	35	0.03	60	< 0.001
20	1.0	40	0.01	99	< 0.001

*Pe = probability that a basecall is erroneous ***(6)***.

Typically good quality values for pure bases should be 20 or higher. Much lower QVs are tolerated for mixed bases, with values >30 being rare ***(6)***. The Q20 rule has been used in many sequencing projects to measure the effective length of a DNA read. A pure base QV <20 will identify the low quality regions ***(8)***.

The KB basecaller, as well as trimming poor quality sequence, allows more correct calls at the 5’ end when compared with the more conventional sequencing analysis software programs ***(6)***.

3. Methods

From the toolbar, click Tools for SeqScape Manager or File to open a New Project (for new projects) or Open Project (for previously set up ones). The SeqScape Manager window is used for setting up most of the parameters, reference sequences and project template. The Project window is used for importing samples for analyzing, data viewing and result generation (**Fig. 1**).

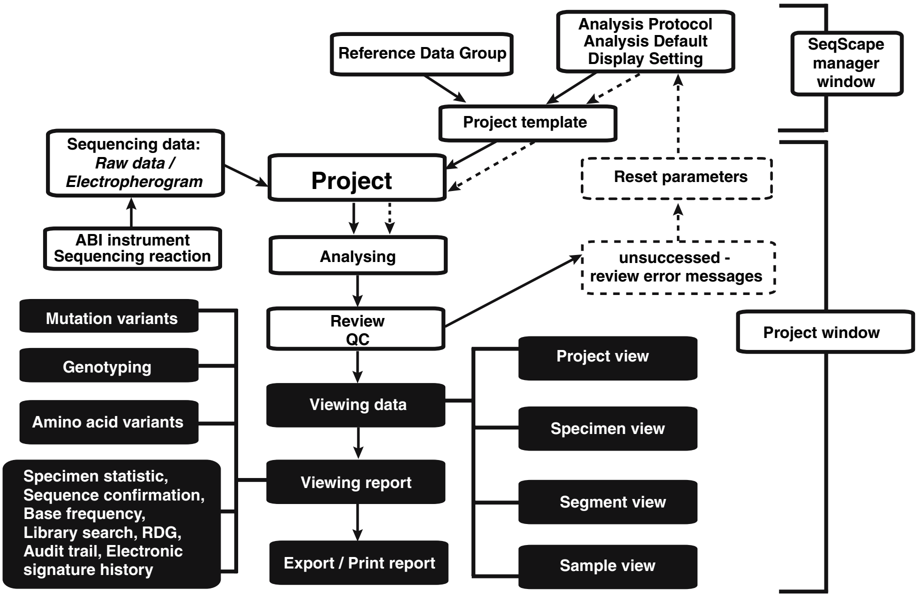
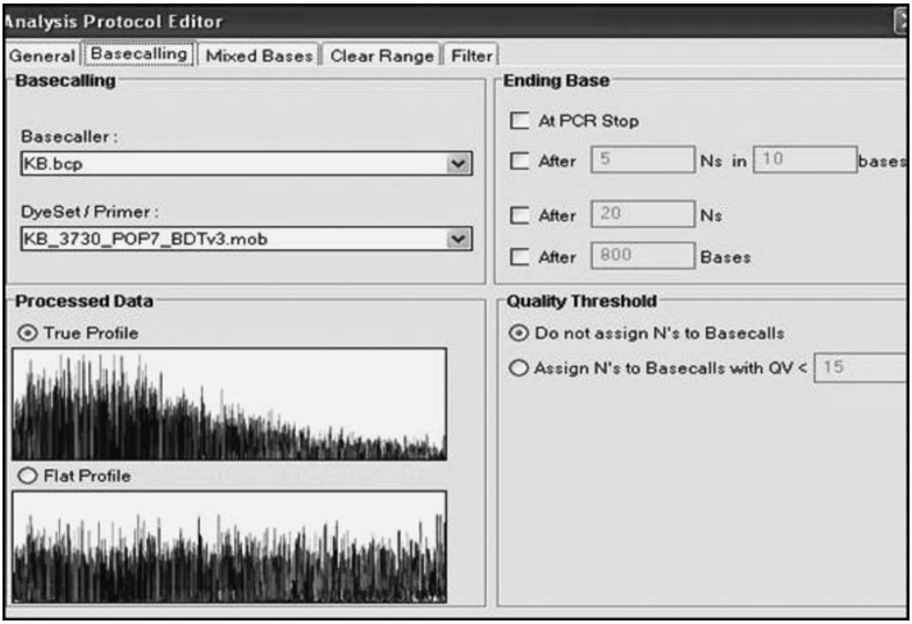


Fig. 1. A summary of Seqscape software procedure. Solid arrows indicate the main steps of analyzing while dotted lines represent steps for re-analyzing.

3.1. Manager Window

3.1.1. Setting up the Parameters

1. Analysis Protocol tab: Open New Analysis Protocol window and enter name. Parameters for Base Calling, Mixed Bases, Clear Range, and Filter are set as in **Fig. 2**. Click the Basecalling tab. Then select from the drop-down menu KB.bcp for Basecaller. If a sequencing reaction was set up using the AB BigDye Terminator Kit (v3.1) and run on the AB 3730 DNA analyzer, the corresponding DyeSet/Primer file on the drop down menu is selected, i.e., KB_3730_POP7_BDTv3.mob file. True Profile is selected for processed data. Ending bases at PCR stop or stopping base calls after a number of ambiguities can be chosen as an option. Either can be used for quality threshold with the QV set at <15. Click Mixed Bases tab. Mixed Bases identification is selected and the IUB (International Union of Biochemistry) code is assigned to the position of any mixed base (**Table 2**) which has the second highest peak ≥ 25% of the main peak. Click Clear Range tab. The Clear Range method can be defined by selecting Use Quality Values and Use Reference Trimming as seen in **Fig. 3**. This will remove bases from both ends until there are fewer than 4 bases out of 20 with QVs ≤ 20. Click Filter tab. The Filter Settings contains



Basecalling

Basecaller: KB.bcp

DyeSet / Primer: KB_3730_POP7_BDTv3.mob

Ending Base

□ At PCR Stop

□ After 5 Ns in 10 bases

□ After 20 Ns

□ After 800 Bases

Processed Data

● True Profile

○ Flat Profile

Quality Threshold

● Do not assign N's to Basecalls

○ Assign N's to Basecalls with QV < 15

Fig. 2. Detail of Analysis Protocol setup.

Table 2
The IUB nucleotide code (*see* http://biocorp.ca/IUB.php).

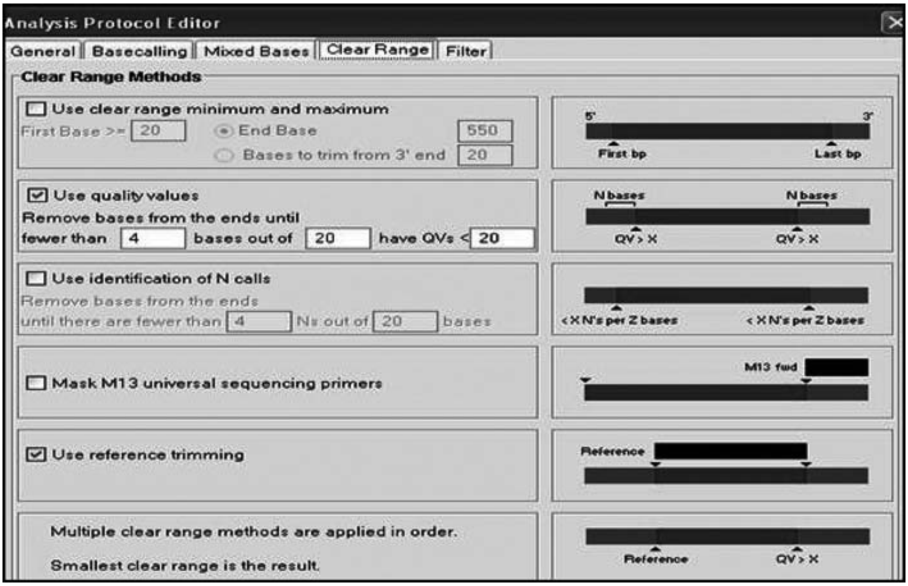
Code	Definition	Mnemonic
A	Adenine	**A**
C	Cytosine	**C**
G	Guanine	**G**
T	Thymine	**T**
R	AG	pu**R**ine
Y	CT	p**Y**rimidine
K	GT	**K**eto
M	AC	a**M**ino
S	GC	**S**trong
W	AT	**W**eak
B	CGT	Not A
D	AGT	Not C
H	ACT	Not G
V	ACG	Not T
N	AGCT	a**N**y

information for rejecting sequences from the assembly. Default values for this setting are recommended (**Fig. 4**).

2. Analysis Defaults tab: Open New Analysis Setting window and enter name. In Project or Specimen tabs there will be two parameters used for setting the gap penalty. They are Gap (opening) Penalty and (gap) Extension Penalty ***(5)***. Their recommended values are shown in **Fig. 5**. An analysis protocol created earlier can be selected from Sample tab and saved.
3. Display Settings tab: Electropherogram characteristics such as color, type, font of bases as well as QV barcode color can be set for easier recognition of errors in base calling. Default values are recommended.

3.1.2. Reference Sequences

1. Obtaining reference sequence(s): This can be obtained directly from GenBank ***(10)*** as a GenBank format (.gb) file (Pane 1, **Fig. 6**) (*see* **Note 2**). Seqscape software also recognizes FastA (.fsta), plain text (.txt) and the sequencing chromatogram (.ab1) files. Reference fragments can be pasted into the program.
2. Reference Data Group tab: A reference sequence entry is required as the input file for the Reference Data Group (RDG) set up. Other steps such as setting regions of interest (ROIs) are arranged through this particular window (Panes 2, 3 of **Fig. 6**).



Clear Range Methods

☐ Use clear range minimum and maximum

First Base >= 20 ● End Base 550

○ Bases to trim from 3' end 20

☑ Use quality values

Remove bases from the ends until fewer than 4 bases out of 20 have QVs < 20

☐ Use identification of N calls

Remove bases from the ends until there are fewer than 4 Ns out of 20 bases

☐ Mask M13 universal sequencing primers

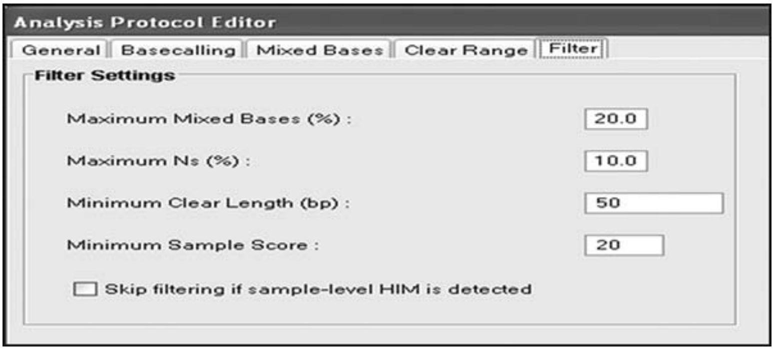
☑ Use reference trimming

Multiple clear range methods are applied in order.

Smallest clear range is the result.

Fig. 3. Setting of Clear Range Methods.

Click New Layer to add a selected ROI. This is graphed as shown in Pane 4 of **Fig. 6**. Information on known nucleotide variants, amino acid variants or a known allele library can be imported or added into the RDG for analysis. Depending on the project plan, more than one reference sequence can be inputted.



Filter Settings

Maximum Mixed Bases (%) : 20.0

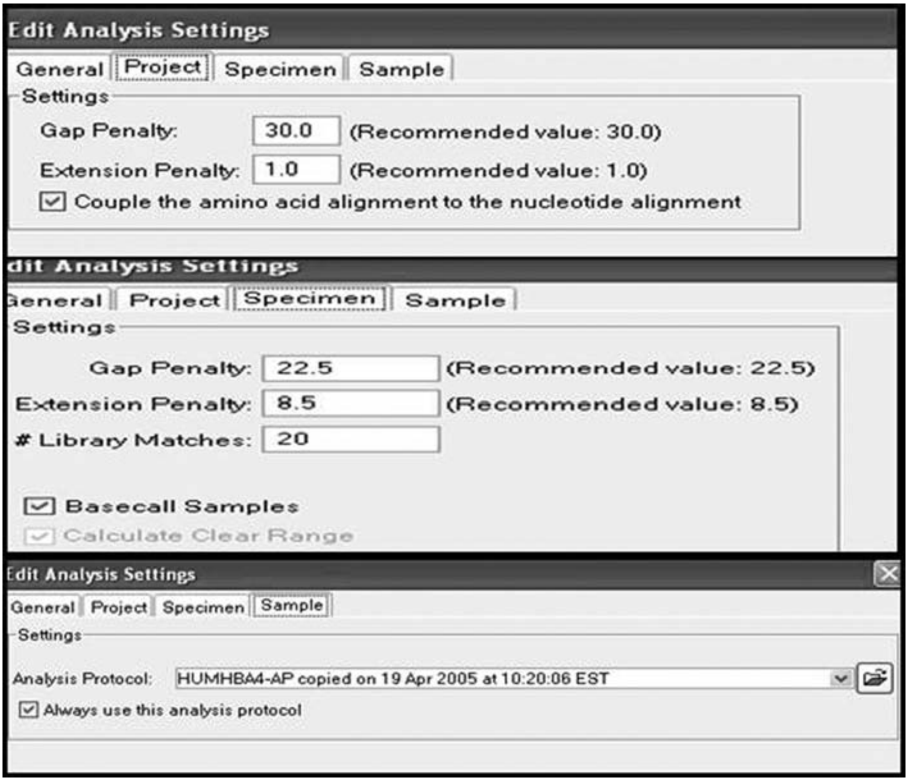
Maximum Ns (%) : 10.0

Minimum Clear Length (bp) : 50

Minimum Sample Score : 20

☐ Skip filtering if sample-level HIM is detected

Fig. 4. Default setting value for filtering process. This value can be changed if necessary e.g. obtaining a pass result from a failed analysis sample.



Settings

Gap Penalty: 30.0 (Recommended value: 30.0)

Extension Penalty: 1.0 (Recommended value: 1.0)

☑ Couple the amino acid alignment to the nucleotide alignment

Settings

Gap Penalty: 22.5 (Recommended value: 22.5)

Extension Penalty: 8.5 (Recommended value: 8.5)

Library Matches: 20

☑ Basecall Samples

☑ Calculate Clear Range

Settings

Analysis Protocol: HUMHBA4-AP copied on 19 Apr 2005 at 10:20:06 EST

☑ Always use this analysis protocol

Fig. 5. Default setting values for gap and extension penalties for project, specimen and sample.

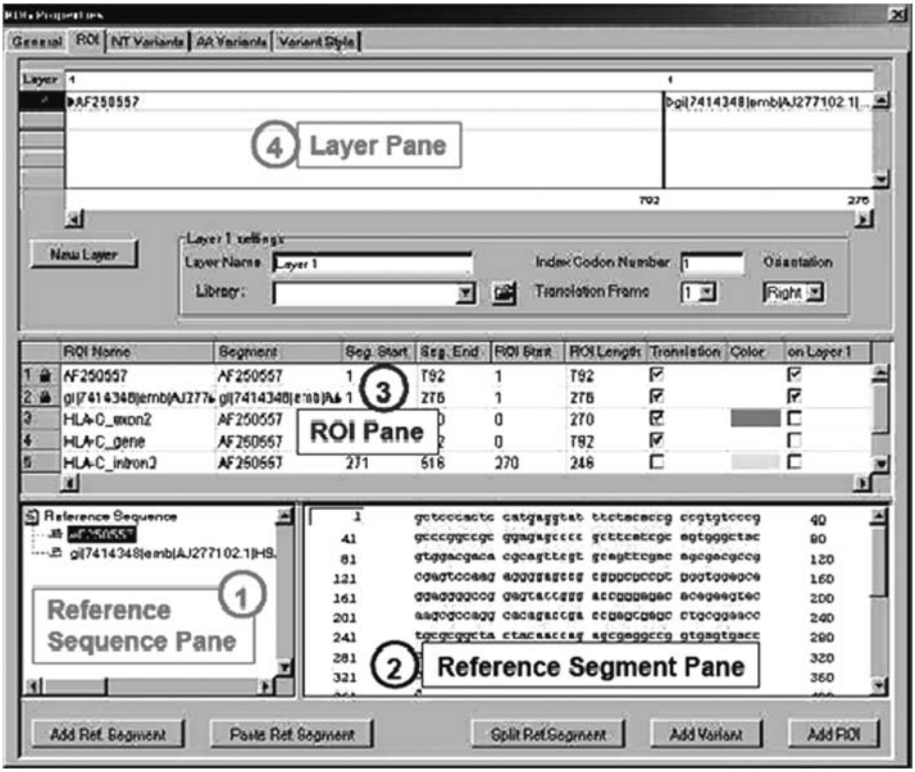
3.1.3. Project Template

After defining analysis parameters and creating the RDG, a project template, which is the collection of the three previous settings (Analysis Defaults, Display Settings and RDG Settings), is set up. The project template can be used as a "common template" and then modified for use with different reference sequences without having to repeat the parameter set up process. Open new project template, enter name and select the three required settings.

3.2. Project Window

3.2.1. Create a New Project

1. Open New Project from file menu: Enter project name then select the appropriate project template containing the analysis strategy which was created previously. A blank project is opened.



Layer Name: Layer 1

Index Codon Number: 1

Translation Frame: 1

Orientation: Right

	ROI Name	Segment	Seg. Start	Seg. End	ROI Start	ROI Length	Translation	Color	on Layer 1
1	AF250557	AF250557	1	792	1	792	☑		☑
2	gi|7414348|emb|AJ277	gi|7414348|emb|AJ	1	276	1	276	☑		☑
3	HLA-C_exon2	AF250557	[illegible]	[illegible]	0	270	☑	▨	□
4	HLA-C_gene	AF250557	[illegible]	[illegible]	0	792	☑		□
5	HLA-C_intron2	AF250557	271	516	270	246	□	▨	□

Reference Sequence
AF250557
gi|7414348|emb|AJ277102.1|HS...

```
  1   gctcccactc catgaggtat ttctacaccg ccgtgtcccg   40
 41   gcccggccgc ggagagcccc gcttcatcgc agtgggctac   80
 81   gtggacgaca cgcagttcgt gcagttcgac agcgacgccg  120
121   cgagtccaag aggggagccg cgggcgccgt gggtggagca  160
161   ggaggggccg gagtactggg accgggagac acagaagtac  200
201   aagcgccagg cacagactga ccgagtgagc ctgcggaacc  240
241   tgcgcggcta ctacaaccag agcgaggccg gtgagtgacc  280
281                                                320
321                                                360
```

Fig. 6. General view of all steps involved in setting up the Reference Data Group.

2. Input samples: Click Import Sample to Project icon on the toolbar to open the Import Sample window. From the drop down menu, search for AB trace data files of interest to import into the project. Samples can be automatically or manually added to the project. Automatic loading of contiguous related fragments is only possible if the samples are given the same prefix. Otherwise, samples are selected manually. For automatic loading go to Start and End and type in the samples required according to instructions on the screen. Highlight required samples and then press Auto Add followed by OK. For manual loading click New Specimen and then highlight sample(s) and click Add Sample followed by OK. A project can have one or more specimens with each specimen comprising various samples. The latter usually represent PCR fragments that make up the entire region being sequenced. A specimen refers to one individual.
3. Data analysis: Once specimens containing samples are imported, a red diagonal line across the specimen name and a green colored ▶ on the menu bar appears. These indicate the sequence needs to be analyzed. Execute the green colored ▶ to start analysis. Following on from the above set up in the project template, SeqScape will

automatically convert raw data image into sequence bases. A QV will be calculated and assigned to each base of the sequence. Mixed bases are identified and assigned by IUB code as described previously. The software will search for sequences with low QV and trimming is carried out to exclude the low quality sequences at both ends of the fragment. Samples with low quality are filtered out, and only good quality sequences remain for the next process of assembly. Specimen consensus sequence is then generated from associated sequences. The next step is to align and compare sample sequences to reference sequences and consensus sequences to reference sequence. The software also automatically searches for matches in the consensus sequence if there is a known variant table or a known allele library table to the project. A summary of the software performance process is shown in **Fig. 7**.

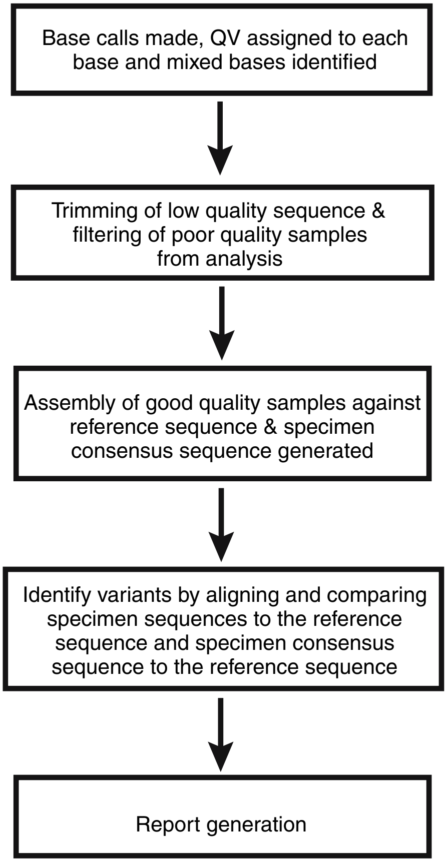

Base calls made, QV assigned to each base and mixed bases identified
Trimming of low quality sequence & filtering of poor quality samples from analysis
Assembly of good quality samples against reference sequence & specimen consensus sequence generated
Identify variants by aligning and comparing specimen sequences to the reference sequence and specimen consensus sequence to the reference sequence
Report generation

Fig. 7. Flow chart of software algorithm.

3.2.2. Open a Previously Analyzed Project

To re-examine a previously completed project that has been analyzed click on Open Project from the file menu and select the relevant file.

3.3. Applications of Seqscape Software

Thalassemias are important blood disorders characterized by a reduced synthesis of one or more the globin chains. Interactions involving the thalassemias, including α- and β-thalassemia can give rise to complex genotypes which manifest as interesting clinical and hematological phenotypes ***(11)***. In these difficult cases, mutation detection by DNA sequencing of a large fragment or a number of overlapping smaller fragments to cover all codon regions, exon–intron boundaries, the promoter and poly A tail regions has become an essential part of the diagnostic work-up. However, after sequencing, manual analysis of data is tedious, time-consuming, and a possible source of error especially when dealing with heterozygotes particularly if sequence quality is marginal. To avoid this we will use the example of α thalassemia to show how rapid analysis of a DNA sequence is possible. An additional benefit is that SeqScape analyzes the data quality and automatically excludes error-prone sequence.

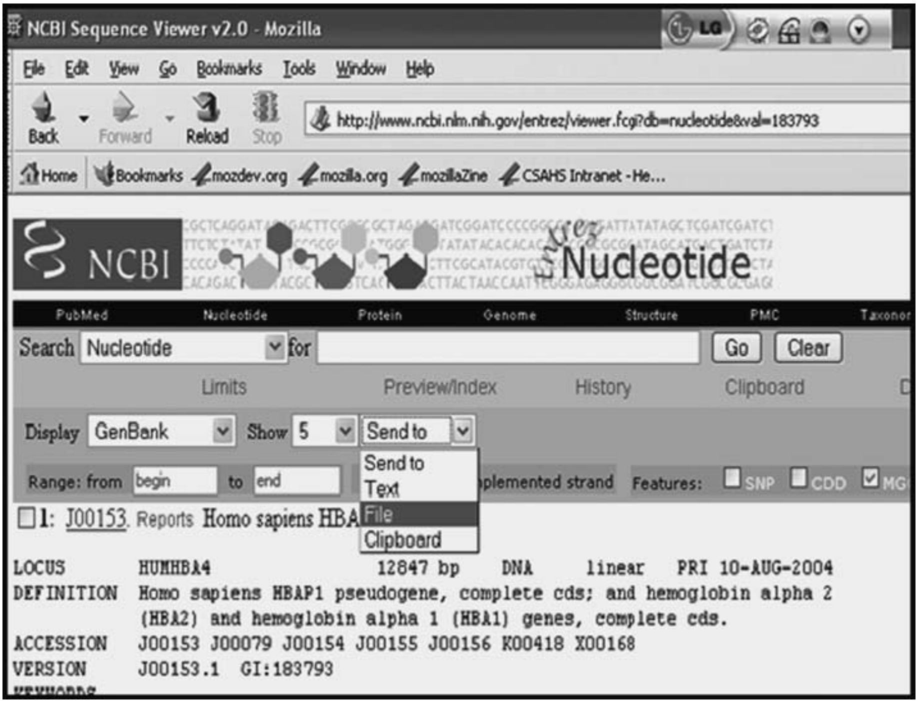
3.3.1. α-Globin Gene Sequence Analysis

Prior to setting up the RDG, reference sequence for the α-globin gene cluster (HUMHBA4) is obtained from GenBank ***(10)***. Find the locus of interest (HUMHBA4) under Search Nucleotide, set Display as GenBank then select Send to File. The setting for downloading a GenBank file is shown in **Fig. 8**. The file is downloaded and then saved under a gb extension, e.g., HUMHBA4.gb, to a specified location ready to be imported into the RDG

3.3.2. The Thalassemia RDG

1. Add reference sequence: Launch SeqScape. Open SeqScape Manager window from Tool menu. Select RDG tab, click New to open the RDG properties. Enter RDG name e.g. HUMHBA4-RDG, then click ROI tab. The following steps allow the user to import the reference sequence.

   Click on Add Ref. Segment tab at the bottom left of the window to open the Import Reference Sequence window. Look for the file of interest, select file name e.g. HUMHBA4.gb, then click Import button. This will import the reference sequence into the Reference Segment Pane. Click OK.
2. Define ROI and Layer: Layers are generated for the purpose of narrowing the analysis stage to a particular region of interest. More than one ROI can be set into one layer but they cannot overlap. From Layer Pane (**Fig. 6**), notice that Layers


NCBI Sequence Viewer v2.0 - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=183793
Home Bookmarks mozdev.org mozilla.org mozillaZine CSAHS Intranet -He...
NCBI
Entrez Nucleotide
PubMed Nucleotide Protein Genome Structure PMC Taxonor
Search Nucleotide for
Go Clear
Limits Preview/Index History Clipboard
Display GenBank Show 5 Send to
Send to
Text
File
Clipboard
Range: from begin to end
plemented strand
Features: □ SNP □ CDD ☑ MG
□ 1: J00153. Reports Homo sapiens HBA

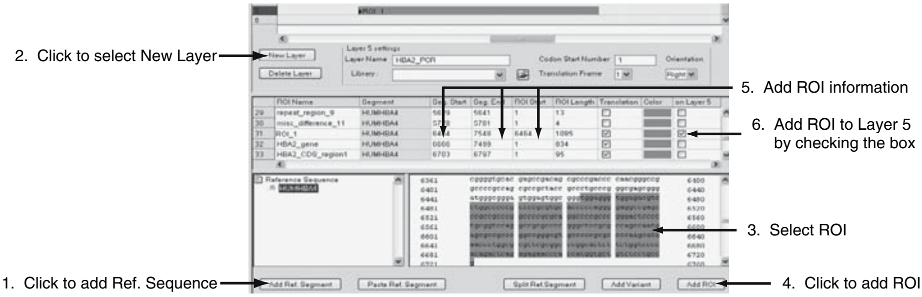
```
LOCUS       HUMHBA4          12847 bp    DNA     linear   PRI 10-AUG-2004
DEFINITION  Homo sapiens HBAP1 pseudogene, complete cds; and hemoglobin alpha 2
            (HBA2) and hemoglobin alpha 1 (HBA1) genes, complete cds.
ACCESSION   J00153 J00079 J00154 J00155 J00156 K00418 X00168
VERSION     J00153.1  GI:183793
```

Fig. 8. Obtaining reference sequence (HUMHBA4) from GenBank.

1–4 are automatically generated by the software. Layer 1 is always the reference sequence and is locked. Layers 2–4 locate the exons of the different genes within the α globin locus, with a different gene in each layer. Click on New Layer to create the next Layer which is Layer 5 and enter Layer Name in the Layer name field, e.g., HBA2-PCR.

In the Reference Segment Pane, select that part of the sequence representing the ROI of interest which has been named HBA2-PCR. This might represent the regions of the α2-globin gene which cover all codon regions, exon–intron boundaries, the promoter, and poly A tail. Note that the nucleotide starting and nucleotide ending numbers highlight the region between these two points. Then click Add ROI. From ROI Pane, the new ROI can be seen as ROI_1. This name can be kept or re-named, e.g., HBA2-PCR as above. Check that the correct information has been added to Segment Start, Segment End for ROI_1, and ROI_1 Start and then tick the check box on Layer 5 column. **Figure 9** illustrates the above setting for the thalassemia RDG.

3. Add known Nucleotide Variants (NT): To add manually, click Add Variant tab. The New NT Variant dialog box will open for manually entering the variant attributes such as variant type, ROI, position (bp), reference base, variant base, description



1. Click to add Ref. Sequence
2. Click to select New Layer
3. Select ROI
4. Click to add ROI
5. Add ROI information
6. Add ROI to Layer 5 by checking the box

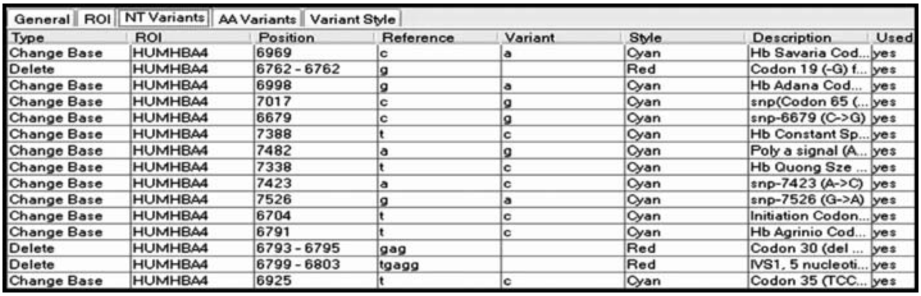
Fig. 9. Steps for setting up the reference sequence/Layer and ROI in the Reference Data Group.

and display style etc. Click OK. It is also possible to import automatically from a table of known nucleotide variants into the RDG properties. A table of known variants can be created using Microsoft Excel. The table display must map to the NT Variant table requirements as seen in **Fig. 10**. The table of variants must be saved in the tab-delimited text file format (e.g., at-variant1.txt) for incorporation into the RDG.

Go to NT Variant tab window and click Import to open the Import NT Variant window. Select the at-variant1.txt under file name field then click Import. An Import Result dialog box appears showing the number of variants imported. Click OK to close the dialog box. Click OK to save the imported variant table.

3.3.3. The Thalassemia Project Template

Click on Project Templates tab, select a previously reported project template and save it under a different name i.e. HUMHBA4-PT. Click OK. Click on



Type	ROI	Position	Reference	Variant	Style	Description	Used
Change Base	HUMHBA4	6969	c	a	Cyan	Hb Savaria Cod...	yes
Delete	HUMHBA4	6762 - 6762	g		Red	Codon 19 (-G) f...	yes
Change Base	HUMHBA4	6998	g	a	Cyan	Hb Adana Cod...	yes
Change Base	HUMHBA4	7017	c	g	Cyan	snp(Codon 65 (...	yes
Change Base	HUMHBA4	6679	c	g	Cyan	snp-6679 (C->G)	yes
Change Base	HUMHBA4	7388	t	c	Cyan	Hb Constant Sp...	yes
Change Base	HUMHBA4	7482	a	g	Cyan	Poly a signal (A...	yes
Change Base	HUMHBA4	7338	t	c	Cyan	Hb Quong Sze ...	yes
Change Base	HUMHBA4	7423	a	c	Cyan	snp-7423 (A->C)	yes
Change Base	HUMHBA4	7526	g	a	Cyan	snp-7526 (G->A)	yes
Change Base	HUMHBA4	6704	t	c	Cyan	Initiation Codon...	yes
Change Base	HUMHBA4	6791	t	c	Cyan	Hb Agrinio Cod...	yes
Delete	HUMHBA4	6793 - 6795	gag		Red	Codon 30 (del ...	yes
Delete	HUMHBA4	6799 - 6803	tgagg		Red	IVS1, 5 nucleoti...	yes
Change Base	HUMHBA4	6925	t	c	Cyan	Codon 35 (TCC...	yes

Fig. 10. A known variant table can be inserted into the RDG for sequence comparison.

Properties tab to open Project Template Editor window then select HUMHBA4-RDG from the drop down menu of Template Elements. Leave the Analysis Defaults and Display Settings the same as before then click OK to save the Project Template. Click close.

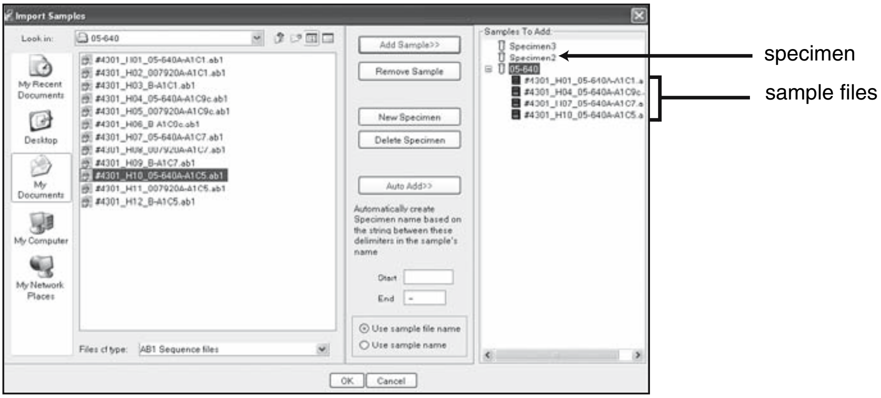
Alternatively one can create a new Project Template and select all required Template Elements as described previously.

3.3.4. Creating the Thalassemia Project Using Project Template

Open a New Project from File, name the project and select the Project Template i.e. HUMHBA4-PT, then click New to open the Project Navigator. Click the Import Sample to Project icon in the toolbar to open Import Sample window. Select folder containing data files that have originated from the sequencing platform e.g. AB 3730 DNA Analyzer. These files are in the form of AB1 Sequence files. Click New Specimen to create a specimen. Enter the specimen name (i.e. 05-640) then select and click Add all related sample data i.e. 05-640-A1C1/ 05-640-A1C7 or all amplicons, into the specimen. Repeat the same procedure for each specimen to be analyzed. At least one additional specimen will be needed – the reference normal sequence for control purposes. Click OK to import specimens and samples into the project as seen in **Fig. 11**.

3.3.5. Thalassemia Data Analysis

The specimen(s) and their related samples are now imported into the project. Click the analysis button ▶ from the toolbar to execute the analysis. The red



Import Samples

Look in: 05-640

#4301_I I01_05-640A-A1C1.ab1
#4301_H02_007920A-A1C1.ab1
#4301_H03_B-A1C1.ab1
#4301_H04_05-640A-A1C9c.ab1
#4301_H05_007920A-A1C9c.ab1
#4301_H06_B A1C0c.ab1
#4301_H07_05-640A-A1C7.ab1
#4301_H08_007920A-A1C7.ab1
#4301_H09_B-A1C7.ab1
#4301_H10_05-640A-A1C5.ab1
#4301_H11_007920A-A1C5.ab1
#4301_H12_B-A1C5.ab1

Files of type: AB1 Sequence files

Automatically create Specimen name based on the string between these delimiters in the sample's name

Start:

End: -

⦿ Use sample file name
○ Use sample name

Samples To Add:

Specimen3
Specimen2
05-640
#4301_H01_05-640A-A1C1.a
#4301_H04_05-640A-A1C9c.
#4301_I 107_05-640A-A1C7.a
#4301_H10_05-640A-A1C5.a

Fig. 11. Importing of thalassemia samples into the project.

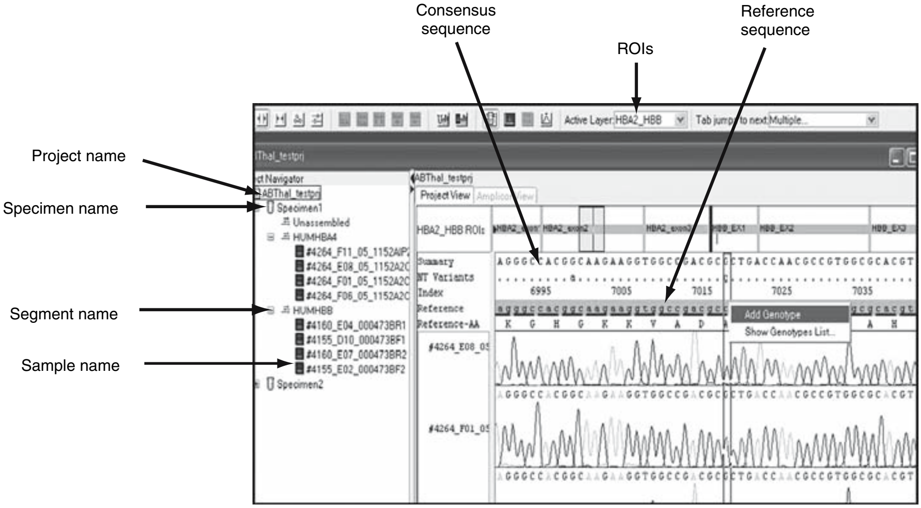


Fig. 12. An example of project view as one of the four different formats for viewing all processed data. Click on Project name to open project view. Other formats such as specimen view, segment view, or sample view can also be seen by clicking to open specimen name, segment, or sample name, etc.

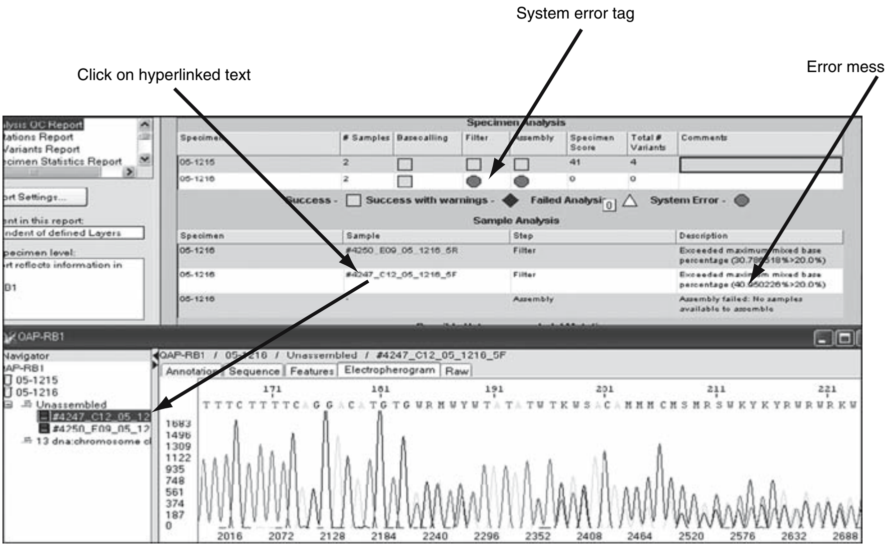
diagonal line across the specimen icon disappears and samples are assembled indicating the analysis process is finished.

From the Active Layer drop down menu select the ROI to view e.g. HBA2-PCR. Processed data can be viewed under four different formats (**Fig. 12**),but before viewing the processed data, one needs to check the analysis performance from the QC report.

3.3.6. QC Report Analysis

Open the Report Manager window under the Analysis icon from the toolbar. The Analysis QC Report is highlighted as the default. This report displays a summary of the project history, particularly the analysis status that indicates how well the system performed. There are four basecalling status indicators that describe the analysis status of specimens. These are: Success; Success with warning; Failed analysis; System error (**Fig. 13**).

To correct this error, click on hyperlink (blue text) of the failed sample (#4250_EO9_05_1216_5R in **Fig. 13**), this will also provide a link to sample data in the project window. Click Electropherogram tab of the sample to view the sequence trace which allows the user to inspect visually and confirm the



Specimen Analysis

Specimen	# Samples	Basecalling	Filter	Assembly	Specimen Score	Total # Variants	Comments
05-1215	2	□	□	□	41	4	
05-1216	2	□	●	●	0	0	

Success - □ Success with warnings - ◆ Failed Analysis - △ System Error - ●

Sample Analysis

Specimen	Sample	Step	Description
05-1216	#4250_E09_05_1216_5R	Filter	Exceeded maximum mixed base percentage (30.795518%>20.0%)
05-1216	#4247_C12_05_1216_5F	Filter	Exceeded maximum mixed base percentage (40.950226%>20.0%)
05-1216	-	Assembly	Assembly failed: No samples available to assemble

Fig. 13. The **top pane** displays the status of the specimen analysis result represented by the basecalling status indicator (System Error). The error message and sample hyperlinked text (when clicked) in the sample analysis part allows the user to investigate or correct sample data in the project view (**bottom pane**).

problem. In this case the sample has failed to assemble due to the filter criteria setting for maximum mixed bases (%) value being set at 20, but the actual percentage of mixed bases was 40.9%. Therefore, no result is generated. Once the problem is resolved, the sample can be re-analyzed and a report will be released.

3.3.7. Re-analyzing Project/Data

Analysis parameters such as incorrect selection of Basecaller or DyeSet/Primer, poor base spacing, incorrect start/stop point selection can all generate suboptimal project results. A heterozygous insertion or deletion mutation (HIM) present in a sample will also produce an analysis failure in the Analysis QC Report. These problems can be improved or corrected by the following two steps, thereby allowing the failed sample to be re-analyzed and passed.

1. Modifying Analysis Parameters in Sample Manager: Open project of interest, select Sample Manager from Analysis menu to open Sample Manager Window. Select the failed sample(s) (indicated by red circle). Depending on error type e.g. incorrect

Basecaller, select the correct Basecaller from the drop down menu, then click OK to go back to the project window. Execute the analysis button ► to re-analyze the sample.

2. Modify Analysis Parameter in an Analysis Protocol: Finding a HIM can also lead to an analysis failure report. In this case, open project of interest, select Sample Manager from Analysis menu of the toolbar. Select sample(s) for re-analysis, click Edit Analysis Protocol , click Filter tab then select the check box to "Skip Filtering if sample HIM is detected". Alternatively, change the default value setting for Maximum Mixed Base (%) from 20 to 50, then click OK. On return to the Sample Manager Window, click Apply, then click OK to go back to the project window. Execute the analysis button ► to re-analyze sample.

3.3.8. Viewing Data

Click Open Project from file menu, select project of interest. Click Open project. Project view can be displayed in many different ways by clicking the following icons (Expanded Nucleotide View, Collapsed Nucleotide View, Expanded Amino Acid View, or Characters/Dots). Processed data are usually analyzed and verified from the project view.

1. Viewing variants in DNA sequence: In Project Navigator**,** click project name to open project view and observe the known (red vertical lines) and unknown variants (blue vertical lines) across the ROI in the top window. Click the ► next to the specimen name to view the electropherogram. Click on any base of the sample consensus sequence then use the tab key to move the cursor from one variant to the next variant. The user can inspect each variant for: 1. QV – Click the QV icon on the main toolbar. The QV is represented by the blue bars on top of each base. Placing the cursor on a blue bar will give its QV. 2. Mixed bases - the IUB code replacement (Table 2) is given, and mixed variants that might be overlooked on the original sequence trace are more easily detectable in SeqScape.
2. Adding variants into the genotype table: Variants found can be added to the Genotype table by right clicking the base for the relevant specimen. Then select Add Genotype. Viewing all data from the project at various levels or formats has facilitated mutation detection. **Figure 14** displays a project view of a heterozygous mutation called Hb Constant Spring (CD142 TAA>CAA) in the α2 globin gene. The mutation was confirmed and viewed by reverse and forward strands and by comparison with a control wild-type DNA sequence.

3.3.9. Viewing Report

Open project of interest, and select project icon. Open Report Manager window from the Analysis menu. Select the report type you wish to view. There are many report formats available. For our purpose, we use the Mutation

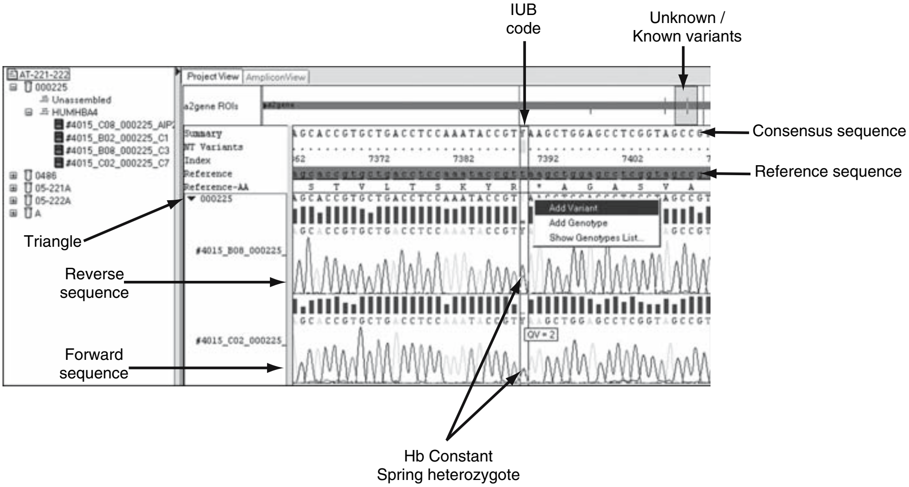


Fig. 14. Variant(s) were detected and shown at the top of the window. Moving the Tab key allows the variant(s) to be viewed for assessment. QVs were used for identifying the mixed bases. The variant detected in this example was consistent with a heterozygous point mutation in the α2 globin gene (Hb Constant Spring). The IUB code here (Y) for the mixed bases indicates C/T. Variants detected can be added to the genotype result table.

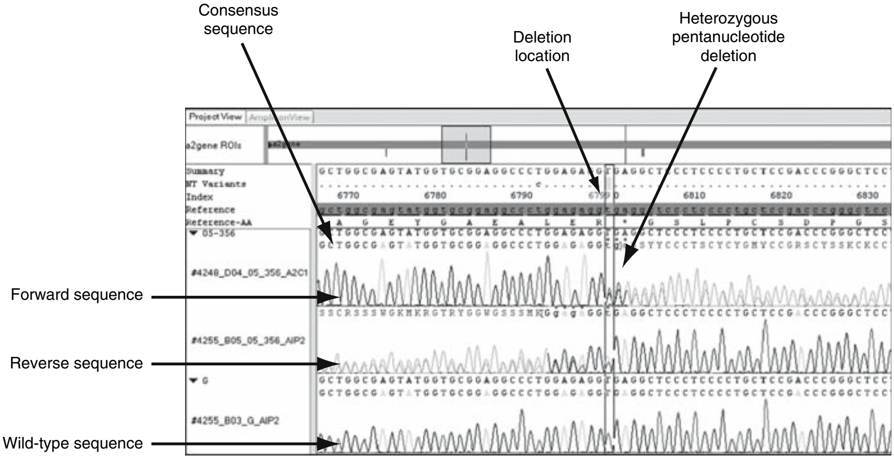


Fig. 15. An example of a heterozygous pentanucleotide deletion in the α2 globin gene. Traces from the first panel (reverse sequence) and the second panel (forward sequence) reveal the heterozygous pentanucleotide deletion compared with the wild-type sequence.

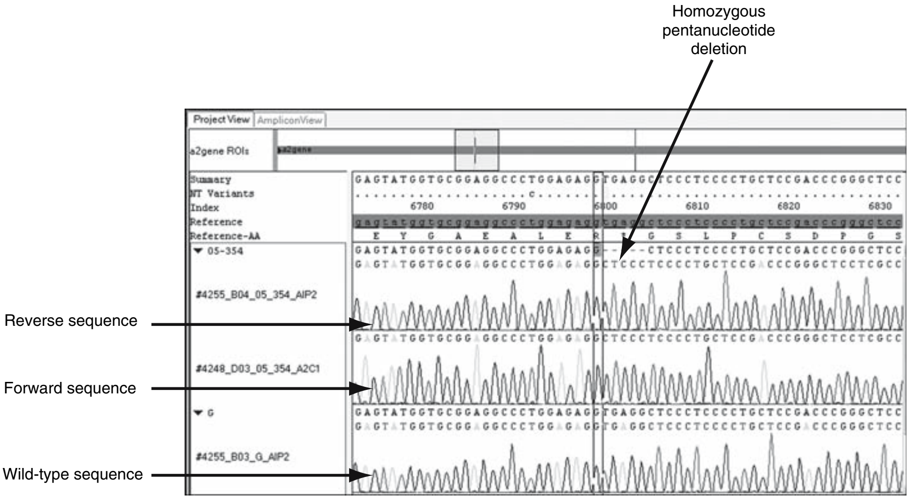

Homozygous pentanucleotide deletion
Project View
AmpliconView
a2gene ROIs
a2gene
Summary
NT Variants
Index
Reference
Reference-AA
05-354
#4255_B04_05_354_AIP2
#4248_D03_05_354_A2C1
G
#4255_B03_G_AIP2
GAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCC
C
6780
6790
6800
6810
6820
6830
gagtatggtgcggaggccctggagaggtgaggctccctcccctgctccgacccgggctcc
E Y G A E A L E R G S L P C S D P G S
GAGTATGGTGCGGAGGCCCTGGAGAGG-----CTCCCTCCCCTGCTCCGACCCGGGCTCC
GAGTATGGTGCGGAGGCCCTGGAGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCC
GAGTATGGTGCGGAGGCCCTGGAGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCC
GAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCC
GAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCC
Reverse sequence
Forward sequence
Wild-type sequence

Fig. 16. An example of a homozygous pentanucleotide deletion in the α2 globin gene. A gap of 5 deleted bases was present in the consensus sequence indicating a homozygous deletion.

report or the Genotyping report. In the Mutation report, the base change is also hyperlinked to data in project view which allows for mutation verification.

Although we have described how to detect missense changes in a DNA sequence, SeqScape is an excellent software package allowing accurate detection of both small deletions and insertions. **Figures 15** and **16** show a

Table 3
An example of a genotype table displaying multiple variants found in the alpha 2 globin gene. The variants' QVs and base locations are shown. Mixed bases are identified with an IUB code.

	Genotype Table					
Specimen	c / 6679/ α2gene	t / 6799/ α2gene	c / 6824/ α2gene	c / 7017/ α2gene	t / 7174/ α2gene	α / 7423/ α2gene
005798	C (46)	T (44)	Y (19)	G (46)	K (11)	C (50)
008055	S (23)	T (42)	C (45)	G (46)	T (42)	C (50)
008057	C (50)	Y (7)	C (23)	G (50)	T (29)	C (50)
008375	C (46)	T (44)	C (42)	G (46)	T (46)	C (32)

heterozygous and a homozygous deletion of TGAGG nucleotides respectively in the α2 globin gene producing another type of α thalassemia.

SNPs and mutations can be identified and any ambiguities quickly resolved. In terms of DNA diagnostic testing, this means greater accuracy and a shorter turnaround time. A Genotyping report result is provided in **Table 3**. A summary of base changes associated with nucleotide positions and their quality values for each sample is displayed.

4. Conclusions

SeqScape software is useful as a high-throughput informatics tool for automated sequence analysis, automated variant identification and for generating reports. It provides accurate and high quality sequence with comprehensive annotation from raw data to result identification in a user-friendly fashion. The analyzed data can be saved, exported, printed and reviewed at any time. The software is expensive to purchase but this is offset by the time saved in analyzing a DNA sequence trace. Further enhancements will allow it to be used in a wider range of DNA based applications (*see* **Note 3**).

Notes

1. Although SeqScape software provides excellent and high-quality analysis of DNA sequence, it cannot do this without prior knowledge of a reference sequence. If the known reference sequence is large (say, about 180 kb) and the computer has inadequate memory or capacity, the program has a tendency to be slow or even freeze. Other limitations of the software are its inability to detect large deletions, and it does not support all sequencing files.
2. For inter-laboratory consistency in reporting, it is necessary to use reference sequences, and the recommended one is RefSeq from the NCBI (*see* **Chapter 11** for more discussion on the importance of reference sequences). Officially approved names for genes can be obtained from the HUGO Gene Nomenclature Committee database (www.gene.ucl.ac.uk/nomenclature/). From the NCBI home page (under Search) select "nucleotide" and type the name for the gene of interest. The search should be limited by selecting RefSeq from the "Only From" in the Limits tag. Accession No. NG_000006 is a 43 kb RefSeq that contains the entire α globin gene cluster on chromosome 16. In this chapter, we have used the smaller non-RefSeq GenBank HUMHBA4 sequence (Accession No. J000153) which contains both α globin genes.
3. Sequencing reagent costs have reduced over the past few years, and the numbers of samples with longer sequence reads have significantly increased along with the increasing number of reference sequences in various databases. More comprehensively integrated software with blast function linked directly to the relevant

databases without the need for entering reference sequence prior to analysis would be a further improvement. More advanced software packages that can call and detect large deletions would also be very beneficial for mutation detection.

References

1. Shendure, J., Mitra, R. D., Varma, C., and Church, G. M. (2004) Advanced Sequencing Technologies: Methods and Goals. *Nat. Rev. Genet.* **5**, 335–344.
2. Davies, H., Dicks, E., Stephens, P., Cox, C., Teague, J., Greenman, C., et al. (2006) High throughput DNA sequence variant detection by conformation sensitive capillary electrophoresis and automated peak comparison. *Genomics* **87**, 427–432.
3. Rieder, M. J., Taylor, S. L., Tobe, V. O., and Nickerson, D. A. (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**, 967–973.
4. Gajer, P., Schatz, M., and Salzberg, S. L. (2004) Automated correction of genome sequence errors. *Nucleic Acids Res.* **32**, 562–569.
5. Alphey, L. (1997) *DNA Sequencing from Experimental Methods to Bioinformatics*. Bios Scientific, Oxford.
6. Applied Biosystems (2004) *SeqScape ® Software Version 2.5 User Guide*. Applied Biosystems, CA.
7. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using Phred.II error probabilities. *Genome Res.* **8**, 186–194.
8. Li, M., Nordborg, M., and Li, L. M. (2004) Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic Acids Res.* **32**, 5183–5191.
9. Ewing, B., Hillier, L., Wendl., M. C., and Green, P. (1998) Base-calling of automated sequencer traces using Phred.I. accuracy assessment. *Genome Res.* **8**, 175–185.
10. http://www.ncbi.nlm.nih.gov/
11. Clark, B. E., and Thein, S. L. (2004) Molecular diagnosis of haemoglobin disorders. *Clin. Lab. Haematol.* **26**, 159–176.

11

Evaluating DNA Sequence Variants of Unknown Biological Significance

Scott A. Grist, Andrew Dubowsky, and Graeme Suthers

Summary

Increasingly, the molecular genetics laboratory has to assess the biological significance of changes (variants) in a DNA sequence. Using the large genes *BRCA1* and *BRCA2* as examples, some approaches used to determine the biological significance of DNA variants are described. These include the characterization of the variant through a review of the literature and the various databases to assess if it has previously been described. Potential difficulties with the various databases that are available are described. Other considerations include the co-inheritance of the variant with other DNA changes, and its evolutionary conservation. Determining the possible effect of the variant on protein function is described in terms of the Grantham assessment as well as identifying functional domains. Studies looking at the distribution of the variant in both the population and the family can also help in assessing its significance. Loss of the variant in a tumor sample would imply that it is not deleterious. Ultimately, it is not any single parameter that helps determine the DNA variants biological significance. Usually this requires multiple lines of evidence.

Key Words: variants, mutations, databases, *BRCA1*, *BRCA2*.

Abbreviations: ESE – exonic splicing enhancer, HGVS – Human Genome Variation Society; OMIM – Online Mendelian Inheritance in Man; LSDB – Locus Specific Data Base; UTR – untranslated region

1. Introduction

With the advent of rapid, low-cost DNA sequencing and other mutation detection strategies in medical laboratories, the primary challenge has shifted

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

from finding variants (*see* **Chapter 10**) to determining if a variant is deleterious. Some genes are highly conserved and any deviation from the accepted "normal sequence" is likely to represent a deleterious variant. However, in outbred populations most genes exhibit a spectrum of variants, ranging from advantageous (reducing the risk of disease), neutral (having no phenotypic significance), low-penetrance (carrying only a modest risk of disease), to deleterious (carrying a high risk of disease). Furthermore, it is clear that the type of variant is not necessarily an accurate guide to whether a variant is deleterious.

In the past, the biological significance of a variant was principally determined by clinical segregation studies and by in vitro functional studies of the variant gene and protein. But the deluge of variants being identified in medical laboratories usually precludes this type of investigative approach. In addition, testing in the medical setting demands a more stringent interpretation than might be required in a research study, and this interpretation must be provided within a limited timeframe.

In this chapter, we provide a framework for the bioinformatics assessment of variants in a diagnostic setting, using the *BRCA1* and *BRCA2* genes as examples. Here we illustrate that in silico investigation has the advantage of not requiring additional laboratory time and capitalizes on the growing resource that variant catalogues provide. However, we caution the reader that utilizing information that has not been verified experimentally also carries risks. While an individual factor may provide compelling evidence in isolation, it is essential that clinical decisions be based on an investigation in which multiple strands of evidence provide a concordant answer.

Throughout this chapter, we refer to the assessment of the "biological significance" of a variant. We use this term to refer to the functional impact of the variant on the usual operation of the gene. This is distinct from an assessment of "clinical significance" or "pathogenicity" which involves an integration of the biological significance and a wide range of clinical information (personal history, histology, family history, etc.). Such an assessment lies well beyond the scope of this chapter. We reserve the term *mutation* for a deleterious variant.

2. Describing a Variant

The unambiguous description used to identify the nature and location of a sequence variant is an essential element in communication between molecular geneticists and health professionals. The reader should be cautioned that there are a number of nomenclature "standards" in existence, some of which are outdated or misused, but are nonetheless still in widespread use. This situation has great potential to introduce errors, both in analysis and clinical testing, and vigilance must be exercised.

Currently, the most accurate and unambiguous nomenclature system gaining international recognition is that endorsed by the Human Genome Variation Society (HGVS). A discussion of this nomenclature system is outside the scope of this chapter, and can be found in detail at the HGVS website (www.hgvs.org). This system provides for description of a variant at the three levels of biological sequence: as genomic DNA, as mRNA (or coding sequence, usually as cDNA), and as peptide. To avoid confusion with other standards, the nomenclature used to describe a variant should always be stated.

The description of a variant must also be precisely located on a specified reference DNA sequence. The usual practice is to use the coding sequence of the gene as the reference sequence, counting the nucleotides from the first nucleotide of the initiation codon. The coding sequence, plus 5'- and 3'-untranslated regions (UTR), are available as cDNA sequences through public reference sequence databases such as those hosted by the US National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) and ENSEMBL (www.ensembl.org). It is important to note that many sequences contain undeclared variants, alternate splicing, missing segments, or errors. By combining sequence data from the human genome project, GenBank, and unpublished data, the NCBI has constructed a series of reference sequences (www.ncbi.nlm.nih.gov/RefSeq/) that have been subjected to curatorial review and provide a higher level of accuracy and consistency than may be found with other sources. In every case, the sequence used to describe a variant should always be stated. Published reports and databases do not necessarily state the nomenclature system or sequence being used to describe variants, and the correct identity of the variant may not be clear.

3. Categorizing a Variant

The first step is to determine the impact of the DNA sequence change on the derived RNA (or cDNA) and protein, describing these changes in the format used by other workers (as detailed above). A full description of the genetic basis of mutations lies beyond the scope of this article, but an overview of variant types is provided in **Box 1**.

The coding sequence of a gene, the exon/intron boundaries, intron sequences, and (if required) genomic sequence can be readily accessed at the NCBI and ENSEMBL sites (among others). For example, the *BRCA1* cDNA (ENSEMBL reference ENST00000309486) can be displayed with separate numbering for

Box 1: Types and effects of DNA sequence variants

Large scale re-arrangements involve disruption of the large scale structure of the gene (deletion, insertion, or inversion of exons or the whole gene) and are usually pathogenic.

Nonsense variants introduce a premature termination codon and result in an mRNA which codes for a truncated protein. Such mRNAs are typically unstable and subject to rapid degradation by the transcriptional machinery. A premature stop codon is generally considered to be pathogenic, but some genes (including *BRCA2*) display alternate or polymorphic stop codons.

Frameshifts typically introduce a stop codon a short distance downstream with the same consequences (and caution) as noted for nonsense variants. However, a frameshift close to the 3’ end of the gene may yield an intact protein with an incorrect COOH-domain.

Missense variants are frequently difficult to interpret because it is not necessarily clear whether the change in peptide sequence affects protein function.

In-frame deletion or insertion yields a protein that is similar in size to the wild-type and it is not necessarily clear if the change is pathogenic.

Silent substitutions in an exon (i.e., same amino acid coded despite the change in codon) are frequently benign polymorphisms. However, some substitutions may activate a cryptic splice site or modulate splicing by affecting an exonic splice enhancer (ESE).

Splice site mutations are usually commonly confined to the critical conserved nucleotides around splice donor and acceptor sites. But it should be noted that variants distant from these splice sites can also affect splicing (as noted above).

Promoter variants may impair transcription or produce an alternate start of transcription. Point mutations can have a profound effect on the promoter efficiency but this can be difficult to predict. Large genomic deletions involving the promoter are likely to be pathogenic.

5’ UTR variants may produce an alternate translation start-site by creating a new AUG upstream of the usual initiation codon.

Intronic variants are usually polymorphisms but may cause mis-splicing or other abnormalities in RNA processing. Conserved intronic regions may encode small interfering RNAs (siRNA) that play a major role in such processing.

3’ UTR variants may also interfere with siRNAs, but this is currently difficult to predict.

the cDNA, coding, and peptide sequences; exon boundaries can be displayed with 25 or more nucleotides at each end of each intron; the display includes common single nucleotide polymorphisms (SNPs). Careful assessment of this display will allow the reader to identify exactly where the variant is located in the functional structure of the gene.

If the variant occurs within the coding region of the gene, examine the cDNA sequence and derived peptide sequence of both the reference sequence and the variant sequence. This may be performed manually, but can be greatly assisted by the use of alignment software, e.g., the utilities freely available at http://au.expasy.org/tools/. Establish the reading frame of the coding sequence and refer to a codon usage table to determine the change to the amino acid sequence predicted from the nucleotide sequence (or use tools such as those listed above). Determine the type of variant (*see* **Box 1**) but do not jump to a conclusion regarding the significance of the variant.

If the variant occurs in a non-coding region, it may be more difficult to determine the biological significance. However, by addressing a systematic series of questions the effect of the change may be revealed: Is the variant located close to or within the gene promoter region? Is the variant predicted to encode an alternate transcription start site? Does the variant introduce a new initiation codon (AUG) in the 5’ UTR? Does the alternate start codon predict a different reading frame? Does the variant disrupt the highly conserved bases in the first or second position at the start or end of an intron, resulting in an alternately spliced transcript? Would such a splice variant result in a frameshift downstream from the variant? Is the variant in the 3’ UTR?

Prediction algorithms are available that can assist in identifying splice donor/acceptor sites (NetGene2: www.cbs.dtu.dk/services/NetGene2/) and exonic splice enhancers (ESE) (ESEfinder: http://rulai.cshl.edu/tools/ESE/). Both normal and variant sequences should be assessed and compared. The output indicates putative candidate ESE and splice sites, usually with an indication of relative efficiency or probability of splicing. It should be remembered that these algorithms are only theoretical predictions and are far from definitive. Nonetheless, they do provide a useful starting point for categorizing such variants. When the mRNA sequence resulting from the splice variant has been determined, consider the following questions: Does the variant mRNA lose or gain coding sequence? Is the correct reading frame retained? Is a new stop codon introduced? Note that a mutation may cause both a frameshift and alter splicing such that the variant splicing restores the reading frame.

Many genes can yield multiple transcripts which are derived from different splicing patterns, each encoding unique functional products. There is usually one predominant transcript that is the basis for assessing variants in the literature. But the reader needs to be mindful that a variant may affect some transcripts but not others arising from the one gene. The ENSEMBL site lists the various transcripts that have been described for a gene in which all are aligned in an overlapping manner to a single genomic framework. This provides an illustration from which an assessment can be made of the potential involvement of a variant in each functional transcript.

4. Obtaining the Evidence

Variant sequences that are predicted to yield a substantially truncated protein i.e. large deletions, nonsense, and frameshift mutations, are usually biologically significant and can be classified as deleterious. However, there are exceptions. Truncating variants in the COOH-domain of *BRCA2* can be neutral polymorphisms ***(1)***. Conversely, a silent variant may not cause any alteration in the predicted protein but can disrupt an ESE resulting in variant splicing in the mRNA ***(2)***. So it is essential that assumptions are not made in interpreting the biological significance of a variant.

On the other hand, variants that retain the reading frame, including missense variants, are much harder to classify. This section presents some principles to assist in the classification of these variants.

4.1. Checking the Literature and Databases

If you can find a detailed assessment of the variant in the biomedical literature or database, this may be the quickest way of determining the biological or clinical significance of a variant. But we caution the reader that both the literature and on-line databases are replete with inconsistencies regarding the significance of gene variants. Before utilizing an interpretation to guide clinical management, ensure that multiple lines of evidence point to the same conclusion.

4.1.1. Searching the Literature

Search the biomedical literature using tools such as PubMed (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), or Google Scholar (http://scholar.google.com/) for papers which describe the variant. These search tools use text matching to identify abstracts and titles of interest, and a successful search demands that you use the same character string as the authors to describe the gene, variant, and location, and that this character string appears in the abstract or title. If different gene names, reference sequences, or nomenclatures are in common use, you may need to search using different descriptions of the variant. You may also need to repeat the searches using genomic, cDNA, and protein descriptions of the variant.

The arbitrariness of variant descriptions in the literature, as well as the possibility that the variant may not be described in the abstract, limits the usefulness of literature searches. However, this disadvantage is offset by a number of key advantages. First, the published assessment may have been peer reviewed (depending on the publication), thereby providing some authority to the assessment. Second, if the details of the assessment are included in the article, the primary data can be reviewed. Third, the authors will have (hopefully) placed the assessment of the variant in a broad context, noting prior references and the assessment of similar variants.

4.1.2. Locus-Specific Databases

There are many on-line databases of variants in individual genes (locus-specific databases, or LSDBs) currently available. These have often been developed as a tool to facilitate research between collaborating laboratories, and address some of the shortcomings of the biomedical literature as a repository for variant reports. In particular, many more variant reports are usually cited in LSDBs than are in the medical literature. But the utility of many LSDBs has been compromised by a lack of consistency in the type of information collected, nomenclature, variant interpretation (even in the same LSDB), curatorial quality and maintenance ***(3)***. To complicate matters further, in the case of many genes there are multiple LSDBs. The one report of a variant may have been replicated in multiple databases, giving the misleading impression that the one interpretation has been endorsed on multiple occasions (*see* **Chapter 12** for further discussion on DNA variant databases).

In addition to LSDBs, there are a number of databases that collate the variants identified from a number of genes using consistent gene names, reference sequences, and nomenclature. With growing recognition of the need for a consistent and comprehensive data repository, this is a welcome move toward having a single central repository of variants that is linked to individual LSDBs ***(4)***. But databases can only be as good as the quantity and quality of the submitted data. Increasing volumes of variant data generated by busy diagnostic laboratories are often not released to LSDBs or a central repository due to time and resource limitations. Hence the reader must be aware that the information held by online databases represents only a subset of the information that is potentially available (*see* **Note 1**).

4.1.3. Central Databases

Three central databases are currently available, each with advantages and limitations.

1. The Cardiff Human Gene Mutation Database (www.hgmd.cf.ac.uk) has the advantage of being more comprehensive than other central databases because the variant reports were sourced from LSDBs, publications, and conference reports. The database is part-funded by a commercial partner, and some resources are only available to subscribers. The accrual of variants in the last decade has fluctuated according to the source and level of funding.
2. Online Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov) is an annotated list of Mendelian traits and genes in humans, with references to relevant literature and other websites. The quality and comprehensiveness of the annotations varies. There are separate entries for genes and traits (diseases), with the gene entries generally being restricted to the biology of the gene. The disease entry includes a section on the underlying molecular genetics (if known), and lists key variants that inform the clinical discussion. Hence OMIM includes a highly selected and idiosyncratic list of variants that are usually regarded as unequivocally deleterious. It does not provide a commentary on variants of uncertain biological or clinical significance.
3. The Waystation (www.centralmutations.org) is a recent endeavor to provide a single point of reference for all human gene variants. It has the advantage of consistency but has a limited volume of data.

4.1.4. Searching for a Variant in LSDBs

The first step is to find the relevant LSDBs. Two of the central databases, the Cardiff Human Gene Mutation Databases and the Waystation, provide links to the LSDBs that were used to source the central database. However, there may be LSDBs for new genes or additional LSDBs for established genes that are not listed. The HGVS also maintains a list of LSDBs for different genes (www.hgvs.org).

Some LSDBs are open to the public. Others require registration via the curator, and some are restricted to certain users. Readers must also bear in mind that LSDBs vary widely in scope, quality, and intent. Some are focused on research issues, while others are designed to assist clinical decision-making. If there are multiple LSDBs for the one gene, it is important to check each LSDB, bearing in mind the possibility that the one report of a variant may be cited in multiple databases.

For example, the largest repository of variant data for the *BRCA1* and *BRCA2* genes is the Breast Information Core (BIC) (http://research.nhgri.nih.gov/bic/). Users must register (no charge) to access the database and can search for a variant using a variety of search parameters. BIC uses the GenBank sequence U14680.1 for *BRCA1* rather than the NCBI-specified reference sequence, i.e., NM_007294.2. It is at this point that an example of the differences between nomenclature systems is clearly demonstrated. BIC nomenclature numbers from the first base of the U14680.1 sequence rather than the usual convention of starting numbering at the first base of the initiation codon (which is at position +120 in U14680.1). There are also differences in the 5' UTR of exon 1:

1. U14680.1 lacks the leading 82 nucleotides at the 5' end of exon 1 present in the NCBI sequence (NM_007294.2).
2. Bases 105-7 and 109 of the NCBI sequence are discordant with the aligned U14680.1 sequence.
3. U14680.1 has an insertion between bases 111 and 112 of the NCBI sequence. This alters the subsequent base numbering in the gene.
4. U14680.1 lacks the entire 3' UTR specified in the NCBI sequence.

With regard to *BRCA2*, BIC uses the GenBank sequence U43746 rather than the NCBI-specified reference sequence, i.e., NM_000059.2. As with the *BRCA1* sequence, BIC numbers the *BRCA2* sequence from the first base rather than using the first nucleotide of the initiation codon, which is at position +229 in U43746. The GenBank and NCBI sequences also differ at a number of non-coding positions in the gene.

1. U43746 has an additional base at the beginning of the 5' UTR of exon 1.
2. The base at 10,854 in U43746 is G rather than A. This is in the 3' UTR.
3. In the 3' UTR, the base at 10,858 in U43746 is not included in the NCBI sequence, thereby altering the numbering for the remainder of the gene.
4. Significant differences exist between the two sequences in the 3' UTR downstream of base 10,926 in sequence U43746.

Further problems arise when a lack of consistency is applied to nomenclature. For example, a deletion of two nucleotides in the *BRCA1* gene which is a common founder mutation in people of Ashkenazi Jewish descent is variously described in the literature as *BRCA1* 185delAG or *BRCA1* 187delAG (both based on the BIC numbering system). The root cause of this inconsistency is that the deletion occurs at a nucleotide sequence of “AGAG” where it is impossible to know whether the first or second “AG” is deleted. In this case, the difference is functionally irrelevant, but unless a standard i.e. assign the change to the most 3' position possible, as recommended by the HGVS, is universally used, difficulties arise when using the nomenclature to search for literature relating to a specific variation. The equivalent HGVS nomenclature for this same mutation is *BRCA1* c.68_69delAG (based on reference sequence

U14680.1) where the lowercase “c” indicates the numbering is based on coding sequence and the more 3’ of the repeated AG nucleotides is deleted.

A conversion may often be required to ensure that user and database curator are addressing the same variant. Information held within BIC lists the number of times that the variant has been reported to BIC, noting the laboratory, publication (if any), and a consensus conclusion regarding its biological significance. The consensus reflects the considered opinion of the curator and colleagues (experts in the field). The great majority of BIC variant reports come from one commercial diagnostic laboratory that provides *BRCA1* and *BRCA2* testing in America. Although this has the advantage of providing a consistent clinical perspective on variant interpretation, it also means that variants detected by methods other than those used by that laboratory, or in ethnic groups outside America, are under-represented in BIC.

4.2. Co-occurrence with a Mutation

An understanding of the clinical and genetic features of the disorder in question can assist in assessing the biological significance of the variant. In the absence of consanguinity (and in an outbred population), it would be unusual to observe homozygosity for a mutation in a patient with an autosomal dominant disorder. Hence homozygosity for the variant would suggest that it is unlikely to be deleterious. Similarly, compound heterozygosity involving a documented mutation and the variant (on different alleles) or co-occurrence of the mutation and the variant on the same allele would argue against the variant being deleterious.

This assessment can gain much greater weight if heterozygous and homozygous (or compound heterozygous) mutations in the gene cause different clinical disorders. For example, constitutional loss-of-function mutations in both alleles of *BRCA1* are lethal in the embryo ***(5)***. Similarly, loss-of-function in both alleles of *BRCA2* is often embryonic lethal, but some homozygous or compound heterozygous mutations cause Fanconi anemia, a condition with characteristic clinical and cytogenetic features ***(6)***. Note however, that this assessment depends on the variant and the mutation being on different alleles i.e. occurring in *trans*, and this may not be easy to prove. For a number of reasons it is unlikely that two mutations would be physically present on the same allele i.e. occurring in *cis*. However, if this were the case, the two mutations would not constitute functional homozygosity and would not result in the homozygous phenotypes. It is also possible that the combination of a mutation and a low penetrance variant may be developmentally tolerated.

These observations can be quantified. As discussed by Goldgar and colleagues *(7)*, the likelihood ratio of observing the variant n times, with k events being in conjunction with a mutation, is

$$\frac{(P_2)^k(1-P_2)^{n-k}}{P_1^k(1-P_1)^{n-k}}$$

where p_1 is the probability that a patient with a neutral variant carries a mutation on the other allele, and p_2 is the probability that a patient carries two mutations (on different alleles). For analyzing the co-occurrence of *BRCA1* variants, Goldgar and colleagues *(7)* set p_1 as 4% (the approximate frequency of *BRCA1* mutations in a large data set) and p_2 as 10^{-4} (reflecting the embryonic lethality of homozygous mutations in *BRCA1*).

This analysis is usually less helpful in assessing variants in genes responsible for autosomal recessive disorders. The presence of a variant and two mutations (on different alleles) would argue against the variant being deleterious. The presence of the variant and one mutation (on different alleles) suggests that the variant may be deleterious but the possibility of their being another cryptic mutation (and the variant being neutral) cannot be excluded. The occurrence of a variant without an associated mutation provides no information regarding the variant's significance. In a further twist to the challenges of interpreting recessive variants, a combination of two or more neutral variants in *cis* can result in the allele being deleterious *(8)*.

These analyses usually rely on access to a large and comprehensive dataset which links mutations, variants of uncertain significance, and neutral variants to individual patients, as well as documenting the associated phenotype. Most variants of uncertain significance are rare, and it is the occasional patient with both a mutation and the variant who provides the essential clue to the significance of the variant.

4.3. Evolutionary Conservation

If the amino acid at a specific point in the peptide sequence is conserved among different species, it suggests that the amino acid plays a key role in the function of that protein. Thus, the nucleotide variant encoding the variant amino acid is likely to be deleterious. Multiple sequence alignments of the same gene from different species can be found in NCBI Homologene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) and the ENSEMBL website.

Alignments may also be performed directly by using Basic Local Alignment Search Tool (BLAST) algorithms *(9)*. This is a widely implemented algorithm which forms the basis of many analytical packages, many of which are open sources, e.g. www.ncbi.nlm.nih.gov/BLAST/. A number of variants of this tool exist, the two most commonly used are BLASTp for alignment of protein sequences and BLASTn for alignment of nucleotide sequences. Query sequence data may be submitted in a number of formats, including plain text, and are used to search for matches against appropriate databases. This can be selected for all species or narrowed to relevant phyla. The resulting matched sequences are aligned relative to the query sequence and ranked in order of conservation (*see* **Chapter 13** for more discussion on BLAST).

Interpretation of these alignments should be viewed with care. A lack of conservation at a site is suggestive that the variant amino acids may be neutral. However, this can be misleading if the variant amino acid has different biophysical properties compared with the amino acids tolerated at that position in other species (*see* next section). Conversely, conservation of the same amino acid suggests that a variant amino acid would represent a deleterious mutation, but this can be misleading if the conservation is due to chance rather than to selection for that amino acid. It is also important to confirm, especially when comparing widely diverged species, that the sequences found by the matching algorithm belong to a biologically relevant protein and are not merely short sections of random homology or similar functional motifs in an unrelated gene. A further confounding factor is that closely related proteins could have different roles in different species, with different regions of the gene being under different selective pressures in the two organisms.

Mathematical models have been developed that allow quantification of these evolutionary differences. The development and use of such models lies beyond the scope of this chapter, and the reader interested in the application of such models to the *BRCA1* gene is referred to ***(10)***.

4.4. Biophysical Consequences

In assessing a missense variant, a key issue is the extent to which the variant amino acid might alter the biophysical attributes of the variant protein. The first element of this assessment is determining the degree to which the variant amino acid differs from the usual amino acid. The substitution of a new amino acid with similar properties to the normal type (termed a conservative substitution) would be less likely to produce a structural or functional change than a substitution involving an amino acid with very different properties (termed a non-conservative change).

There are a number of indices for amino acid properties in use, but one of the earliest described and most widely applied is the Grantham score derived from volume (V), polarity (P) and side chain composition (C) of the amino acid ***(11)***. Each component parameter is a continuous variable, with side-chain composition being the ratio of the atomic weight ratio of non-carbon elements in end groups or rings to carbons in the side chain. The three parameters are combined in a formula that yields the best fit with the relative frequencies with which amino acids can substitute for each other in multiple proteins in different species. The pairwise differences in Grantham score are listed in **Table 1**. The greater the deviation from 0, the greater is the biophysical "distance" between the two amino acids.

But it is also essential to know what degree of difference is tolerated at this specific point in this specific gene. At a specified point in an alignment of multiple reference peptide sequences from different species, each amino acid is plotted as a point in three-dimensional space (corresponding to the three components of the Grantham Score). The length of the longest diagonal of the box bounding these points is a measure of the variability tolerated at this point, and is termed the Grantham *Variation* ***(12)***. The distance between the point plotted for a variant amino acid and the closest surface of the box bounding the reference points is termed the Grantham *Deviation*. The quantification of this assessment allows the measure to be incorporated with other parameters to provide a numerical measure of the probability that a variant is deleterious ***(12)***.

4.5. Functional Consequences

There is increasing information available about the presence of specific functional domains in proteins. Domains are defined on the basis of conserved amino acid sequence in different proteins and species, and usually represent the key functional components of elements of a protein. Hence a variant which results in the disruption of a domain, either due to an in-frame deletion or to the presence of a variant amino acid that is markedly different from the usual amino acid, would normally be considered to have a high likelihood of producing a structural or functional change in the mature peptide and is therefore likely to be deleterious. Knowledge of the biological function ascribed to the domain carrying the variant can also assist in deciding what type of further supporting evidence or laboratory testing might be sought to help confirm the bioinformatics prediction.

The catalogue of domains, and the functions associated with them, is growing. Lists of protein domains are available at a number of sites, including the NCBI (www.ncbi.nlm.nih.gov) and ENSEMBL (www.ensembl.org). For example, the ENSEMBL entry for *BRCA1* (reference ENSG00000012048)

Table 1
Grantham differences for each possible amino acid substitution. The greater the deviation from 0, the greater is the biophysical "distance" between the two amino acids

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

includes links to the SwissProt database which lists the relevant domains and locations on the peptide sequence derived from the *BRCA1* gene.

This type of analysis is, by definition, limited to what has been previously observed and entered into the sequence databases. Not all functional domains are mapped and the individual roles of the regions between the functional domains are not fully understood. Computer predictions of protein folding based directly on an input sequence may be one way around this limitation, and indeed a number of algorithms for the prediction of protein folding are becoming available, e.g., http://swissmodel.expasy.org//SWISS-MODEL.html and http://predictor.scripps.edu/. Programs such as these can provide useful insight into the likely effects of amino acid changes, but in the main these computations are still highly developmental and should be regarded cautiously at this time (*see* **Chapter 13** for further information on predicting protein function).

4.6. Occurrence in Normal Populations

A deleterious variant would be expected to occur more frequently in affected patients than in unaffected people. If the variant is found in unaffected populations, this can provide clear evidence that a variant is not deleterious. However, this information is rarely available. Variants of uncertain significance are usually rare and the failure to identify the variation in an unaffected sample provides no information. If the disorder has incomplete penetrance (for whatever reason), the presence of the variant in unaffected relatives also provides little information. For the same reason, population-based studies may identify the variant in a proportion of unaffected people, albeit at a lower frequency than among affected patients.

In assessing information about the frequency of the variant in unaffected people, it is important to assess the statistical significance of population figures that are quoted and the nature of the "control" population. For example, three founder mutations in *BRCA1* and *BRCA2* occur with a combined carrier frequency of more than 2%, and with individual frequencies as high as 1.5%, in unaffected members of the Ashkenazi Jewish community ***(13,14)***. Similar frequencies of mutations for other disorders have also been documented in other ethnic groups. Hence the presence of the variant in the unaffected population is not necessarily a robust indication that the variant is neutral.

LSDBs are a potential source of information about the frequency of variants in an unaffected sample. There are also genome-wide databases that can provide this information, and some of these data are usefully collated for each gene at the ENSEMBL site (www.ensembl.org).

4.7. Segregation in Affected Families

In a family with multiple affected members, the causative mutation would be expected to segregate with disease. Co-segregation suggests that the variant is deleterious, while failure to co-segregate would suggest that it is neutral. Within a small kindred (and assuming the disease (phenotype) and variant are rare), it is simple to calculate the probability of a variant and the disease co-segregating by chance. For example, if two first-degree relatives of the proband have the same dominant disease and variant as the proband, the chance of them having both disease and variant by chance is $(½)^2$. By combining data from multiple families, the likelihood of co-segregation occurring by chance alone can decrease very rapidly and provide strong evidence in favor of the variant being associated with the disease.

However, there are a number of important cautions. First, in diseases with incomplete penetrance, the presence of the variant in an unaffected relative does not necessarily indicate a lack of co-segregation with disease. In such cases, it is best to limit the assessment to the segregation of the variant among affected relatives. Second, some common diseases (such as breast cancer) do also occur as a chance event in kindreds with familial disease. These phenocopies will confound a segregation study. Third, if the variant is common, two relatives may have the same variant without this being inherited from a recent common ancestor.

For these reasons, it is often preferable to utilize a Bayesian analysis that incorporates these possibilities. This discussion goes beyond the scope of this chapter, but a general statistical approach has been proposed by Thompson and colleagues ***(15)*** and can be readily implemented in the linkage analysis program, LINKAGE ***(16)***.

It is important to remember that a segregation study in one family (or among related families) cannot prove a causal association between variant and disease. It is also possible that the underlying mutation is elsewhere in the gene and the variant is merely linked to it. This likelihood is greatly reduced if the segregation study incorporates unrelated families.

4.8. Loss of Normal Allele in Tumor

In interpreting reports, LSDB entries, and segregation studies when assessing a variant, there is an aspect of cancer biology that can provide useful information. Most familial cancer syndromes are due to inheritance of a loss-of-function mutation in one allele of a tumor suppressor gene. The principle underlying tumorigenesis in these disorders is that a cancer then arises because of a somatic mutation resulting in loss of function of the remaining normal allele. This principle provides an opportunity to assess the significance of a variant in a tumor suppressor gene. If the allele with the variant is retained in the tumor, this provides some evidence that the variant is deleterious. Loss of the variant allele in a tumor would argue strongly against the variant being deleterious. If the variant can be studied in multiple cancers from one or more patients with the variant and is shown to be consistently retained, a relatively small number of analyses can push up the odds in favor of the variant being deleterious quite rapidly ***(17)***. The main caveats to this analysis are that the variant allele may be retained by chance alone and that the variant allele being tested may in fact only be linked with the actual mutation and not deleterious *per se*. Such analyses can also be integrated into segregation studies of a variant.

5. Weighing the Evidence

As emphasized at the outset, evaluation of the biological significance of a variant in a clinical setting requires a more stringent approach than may be required in other settings. For this reason, it is essential that a conclusion be based on multiple lines of evidence. Ideally the various lines of evidence will be congruent, with each item either pointing to the same conclusion or (at least) providing limited evidence either way.

At a number of points in the preceding discussion, we have highlighted that various measures of significance can be quantified. This paves the way for combining the various assessments and calculating the odds that the variant under consideration is causally associated with the disease. This approach has been successfully applied to variants in the *BRCA1* gene ***(7,10,12,17)***. The consensus in these publications has been that odds of 1,000:1 in favor of a causal relationship are sufficient to categorize the variant as deleterious; odds of 100:1 against a causal relationship are sufficient to categorize the variant as neutral; the variant is otherwise classified as being of uncertain biological significance.

We strongly support this approach as it promotes an objective and quantified process for the evaluation of variants, while explicitly recognizing that the assessment is based on limited data. As more data are accumulated, the odds for or against causality will change and it is possible that the biological significance of a variant will need to be re-considered.

There may be additional clinical or pathological data that can inform decisions regarding the significance of a variant. The assessment of such data goes beyond the scope of this chapter, but we caution the reader that each component of the analysis should be explicit, objective, and conservative. The decision regarding the significance of a variant may be translated into medical decision-making with irreversible consequences. In the case of *BRCA1* analyses, the conclusion that a variant is deleterious may lead to presymptomatic testing of relatives and consideration of prophylactic surgery.

6. Note

1. The fact that a variant has been reported elsewhere does not relieve the reader of the need for due diligence in assessing a variant. Always verify the reported interpretation by checking citations and satisfying yourself of its scientific veracity. In cases of uncertainty, it is frequently valuable to communicate directly with the reporting laboratory. It is not unknown for the interpretation of a variant to have been changed and not immediately updated on the LSDB. Also note that some LSDBs focus on heritable (germline) variants while others deal with somatic variants. The interpretation of a somatic variant does not necessarily apply to the germline equivalent.

References

1. Mazoyer, S., Dunning, A. M., Serova, O., Dearden, J., Puget, N., Healey, C. S., et al. (1996) A polymorphic STOP in BRCA2. *Nat. Genet.* **14**, 253–254.
2. Liu, W., Qian, C., and Francke, U. (1997) Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. *Nat. Genet.* **6**, 328–329.
3. Claustres, M., Horaitis, O., Vanevski, M., and Cotton, R. G. (2002) Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res.* **12**, 680–688.
4. Horaitis, O., and Cotton, R. G. (2004) The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum. Mutat.* **23**, 447–452.
5. Lane, T. F., Lin, C., Brown, M. A, Solomon, E., and Leder, P. (2000) Gene replacement with the human *BRCA1* locus: tissue specific expression and rescue of embryonic lethality in mice. *Oncogene* **19**, 4085–4090.
6. Howlett, N. G., Taniguchi, T., Olson, S., Cox, B., Waisfisz, Q., De Die-Smulders, C., et al. (2002) Biallelic inactivation of *BRCA2* in *Fanconi anemia*. *Science* **297**, 606–609.
7. Goldgar, D. E., Easton, D. F., Deffenbaugh, A. M., Monteiro, A. N., Tavtigian, S. V., Couch, F. J., et al. (2004) Integrated evaluation of DNA sequence variants of unknown clinical significance: application to *BRCA1* and *BRCA2*. *Am. J. Hum. Genet.* **75**, 535–544.
8. Harvey, J. S., Carey, W. F., and Morris, C. P. (1998) Importance of the glycosylation and polyadenylation variants in metachromatic leukodystrophy pseudodeficiency phenotype. *Hum. Mol. Genet.* **7**, 1215–1219.
9. 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.

10. Abkevich, V., Zharkikh, A., Deffenbaugh, A. M., Frank, D., Chen, Y., Shattuck D., et al. (2004) Analysis of missense variation in human *BRCA1* in the context of interspecific sequence variation. *J. Med. Genet.* **41**, 492–507.
11. Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864.
12. Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., et al. (2006) Comprehensive statistical study of 452 *BRCA1* missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43**, 295–305.
13. Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., et al. (1997) Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast–ovarian cancer families. *Am. J. Hum. Genet.* **60**, 1013–1020.
14. Roa, B. B., Boyd, A. A., Volcik, K., and Richards, C. S. (1996) Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. *Nat. Genet.* **14**, 185–187.
15. Thompson, D., Easton, D. F., and Goldgar, D. E. (2003) A full-likelihood method for the valuation of causality of sequence variants from family data. *Am. J. Hum. Genet.* **73**, 652–655.
16. Lathrop, G. M., Lalouel, J. -M., Julier, C., and Ott, J. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. U S A* **81**, 3443–3446.
17. Chenevix-Trench, G., Healey, S., Lakhani, S., Waring, P., Cummings, M., Brinkworth, R., et al. (2006) kConFab Investigators. Genetic and histopathologic evaluation of *BRCA1* and *BRCA2* DNA sequence variants of unknown clinical significance. *Cancer Res.* **66**, 2019–2027.

12

Developing a DNA Variant Database

David C. Y. Fung

Summary

Disease- and locus-specific variant databases have been a valuable resource to clinical and research geneticists. With the recent rapid developments in technologies, the number of DNA variants detected in a typical molecular genetics laboratory easily exceeds 1,000. To keep track of the growing inventory of DNA variants, many laboratories employ information technology to store the data as well as distributing the data and its associated information to clinicians and researchers via the Web. While it is a valuable resource, the hosting of a web-accessible database requires collaboration between bioinformaticians and biologists and careful planning to ensure its usability and availability. In this chapter, a series of tutorials on building a local DNA variant database out of a sample dataset will be provided. However, this tutorial will not include programming details on building a web interface and on constructing the web application necessary for web hosting. Instead, an introduction to the two commonly used methods for hosting web-accessible variant databases will be described. Apart from the tutorials, this chapter will also consider the resources and planning required for making a variant database project successful.

Key Words: DNA variant, database, domain knowledge, database project planning.

Abbreviations: cDNA – copy or complementary DNA; CGI – common gateway interface; CRC – class responsibility collaboration; DDL – data definition language; gDNA – genomic DNA; DBMS – database management system; HGVS – Human Genome Variation Society; HTML –hypertext text markup language; HUGO – Human Genome Organization; SQL – structured query language; OMIM – Online Mendelian Inheritance in Man

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

Disclaimer: The sample data listed in **Fig. 8** are fictitious and should only be used for educational purposes. They should not be used for studying the genetic epidemiology of either cutaneous malignant melanoma or pancreatic carcinoma or for diagnostic purposes

1. Introduction

Variant databases are proving to be an increasingly valuable information resource in both the clinical and research genetic environments. At the time of writing, the total number of published disease-associated variants curated by the Human Gene Mutation Database, Cardiff (HGMD®) exceeded 60,000 *(1)*. To keep track of the growing inventory of DNA variants, many laboratories employ information technology to store and to distribute the data and its associated information to clinicians and researchers *via* the Web. To date, there are as many as 970 locus-specific, 8 disease-centered, 4 mitochondrial, and 19 central variant databases listed on the Human Genome Variation Society (HGVS) web site *(2)*. Locus- and disease-specific variant databases co-exist in a symbiotic relationship with the central databases. Because of the expert knowledge contained in a specific range of loci or diseases, locus- and disease-specific variant databases built by specialists' laboratories often add value to the central databases. On the other hand, the user volume of central databases is much larger than that of any individual variant database. Hence, specialist databases rely on their association with the central databases to increase visibility to the user community.

The hosting of a web-accessible database requires collaboration between bioinformaticians and biologists and careful planning to ensure its usability and availability. In this chapter, the reader will learn about the resources required for hosting a web accessible variant database.

2. Methods

The process of developing a DNA variant database can be divided into four stages: inception, elaboration, construction, and transition *(3)* (*see* **Note 1**).

1. Inception: The stage when the functionality of the database and scope of the data curation are being defined. It is also the stage when the availability of human, financial, and computing resources is being assessed. If the database is going to be developed by a third-party consultancy, its functionality must be clearly defined by the project leader at this stage. Most project failures can be attributed to ill-defined functional requirements.
2. Elaboration: The stage when operational scenarios are being modeled as use cases. The purpose is to model possible interactions between users and the database.

A suitable software design for the web application will also be decided at this stage (*see* **Note 2).**

3. Construction: The stage when use cases are being converted to a series of engineering diagrams which will collectively serve as a guide to the programming process. Programming will be done at this stage.
4. Transition: The stage when the service is being made available to users. Activities include software testing, code optimization, and deployment.

2.1. Inception

2.1.1. Human Resources Requirement

There are three roles required for the planning and building of a DNA variant database: project leader, programmer, and curator. The project leader is the person who initiates the project. His/her tasks include assessing the feasibility and practicality of the database project, budgeting for the project, and recruiting one or more programmers and curators. The project leader is often the chief investigator for the research project. The programmer is the person who designs and implements the database. His/her task includes assessing the availability of computing and human resources, and resolving any technical issues encountered during the construction and transition stages of the project. The curator is the person who collects the data on DNA variation and populates the database. In some laboratories, the same person may play the dual roles of a programmer and a curator.

2.1.2. Computing Resource Requirement

2.1.2.1. Hardware and Software.

Unless one is attempting to build a central database like the HGMD®, locus-specific and disease-centre variant databases tend to store less than 2 megabytes of data. As such, a workstation equipped with the latest Pentium processor should be sufficient for a typical database project. Because a database is basically an implementation of a data structure, software that can automatically build a database according to a pre-defined schema and can read and write data into the database is required. This type of software is known as a database management system (DBMS). Many DNA variant databases in service have been implemented on either of the two open-source DBMSes, i.e., PostgreSQL or MySQL. Enterprise-level commercial DBMSes, e.g., Oracle® and Sybase®, are often used for implementing institutional data warehouses that store not only DNA variation data but also genome-wide sequence data, microarray data, protein- protein interaction data, and even gene and medical ontologies.

Unless an on-site installation is already available to the reader, the size of a locus-specific or disease-centre DNA variant database is usually too small to call for the service of an enterprise-level DBMS. For the purpose of writing a database schema, a text editor is also required.

2.1.2.2. Service Provider

A more important issue to decide on is who will be the service provider? This depends on the operating environment of the organization within which the project leader is working, the availability of skilled personnel, and the budget allocation for the project.

2.1.2.2.1. In-House Hosting. The biggest advantage of in-house hosting is that the curator has full control of data auditing and web hosting. The curator can make the latest data accessible on the web soon after a novel variant has been detected or published. However, this approach requires a curator skilled in both database administration and web hosting. Therefore in-house hosting is most suitable for laboratories that have the personnel with the appropriate skill available or a research institute equipped with an in-house bioinformatics team. In the latter case, the project leader will need to negotiate with the bioinformaticians on the separation of roles i.e. the data curation part will be handled by the appointed laboratory staff whereas information technology-related issues, such as data security, database administration and web hosting, will be the responsibilities of the bioinformatics team. If it is a research laboratory, e.g., a university teaching hospital research group, the project leader may have to consult with the institution's information technology service.

2.1.2.2.2. External Provider. The alternative to in-house hosting is to assign the task to a commercial internet service provider or to a not-for-profit organization such as the HGVS. Unless they are in the business of providing bioinformatics services, few commercial providers have the experience of hosting web accessible databases for scientific use. They usually charge a regular fee for their service. Hence, readers are recommended to submit their data to the WayStation/Central Database administered by the HGVS if their laboratories do not have the resources to host their own database ***(2)***.

2.2. Elaboration

2.2.1. Functional Requirements

The functional requirements of a database are the written description of its functionality. As shown in **Fig. 1**, the functional requirements of a variant

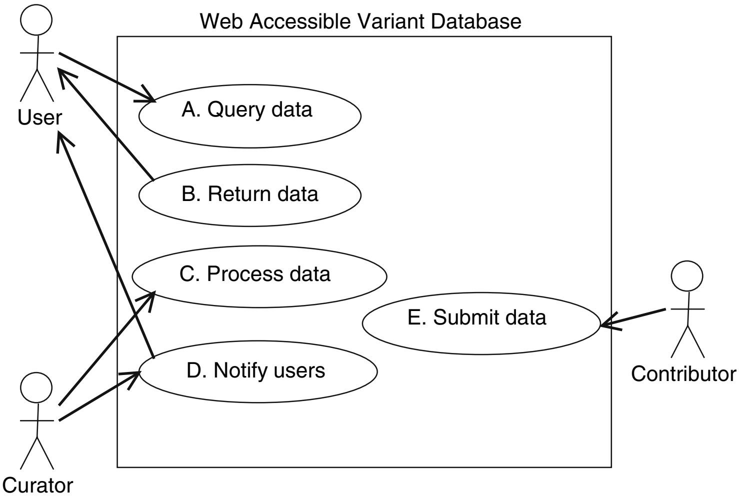


Fig. 1. Use case diagram depicting the functional requirements of a typical variant database.

database include (1) querying data through a web interface, (2) returning queried data to the web interface, (3) processing the curated data, (4) notifying users, and (5) submitting data. While it may not be necessary to have the functions of steps 4 and 5 automated, the provision of functions 1–3 is essential.

2.2.2. Data Requirements

2.2.2.1. Scope of Data Curation

According to the HGVS recommendations, a variant database should store as minimal core data the gene symbols of the disease loci, gDNA variants, RNA or cDNA changes, residue changes, and citations. Variations should be curated in the form of HUGO nomenclatures. The simplest form is single base substitution. For example, a gDNA change of guanosine to adenosine at nucleotide position 1,000 of a disease locus should be annotated as g.1,000G>A. For more details about HGVS-recommended nomenclatures, the reader is referred to ***(4)*** and ***(5)***. Auxiliary data are supplementary to the core data. These include geographical distribution, population occurrence, detection method, kindred, and citation. There is no minimum requirement as to how detailed auxiliary data should be curated. It depends on the complexity of the disease and the extensiveness of the data curation.

2.3. Construction

2.3.1. Mapping Domain Knowledge to Database Schema

The definition of the term *domain knowledge* in the current context means knowledge about human genetic variations. There are two methods for mapping knowledge to database schema, namely fine-grain mapping and coarse-grain mapping. In the former, every datum would be mapped to a unique attribute of a certain class (*see* **Note 3**). In a relational database, data are being stored in a tabular form with each class being a table. It was called fine grain because the value of every attribute is atomic. In coarse-grain mapping, every set of contextually related data would be mapped to an attribute. Thus the value of each attribute can be a composite of several pieces of data ***(6)***. The decision as to which method to be used would depend on a number of factors:

1. Complexity of the disease which can be viewed as

    a. *Genetic heterogeneity*: Is the disease monogenic or multigenic for the same phenotype? For example, hereditary cardiomyopathies ***(7,8)*** and familial malignancies ***(9)*** are often multigenic. Classic examples of monogenic diseases are cystic fibrosis and hereditary fructose intolerance.
    b. *Clinical heterogeneity*: Does the locus variant associate with more than one disease phenotype for the same spectrum? For example, *CACNA1A* gene variants have been found in patients afflicted with episodic ataxia and also in those with hemiplegic migraine ***(10)***. Do carriers of the same variant but of different ethnicities differ in their clinical manifestation such as a disease onset age or mortality rate within a particular age group? Does the locus variant associate with cross-spectrum phenotypes? For example, *SCN5A* gene variants have been known to associate causally with the Long QT Syndrome but may also associate with gastrointestinal ailments ***(11)***.
    c. *Occurrence*: Is the locus variant found in diverse ethnicities? For example, variants associated with familial hypertrophic cardiomyopathy have been found in Asian ***(12,13)*** and Caucasian populations ***(7,8)***, but variants associated with cystic fibrosis have been found predominantly in the Caucasian population ***(14)***.

    Generally speaking, coarse-grain mapping is most suited to locus-specific databases of monogenic diseases that manifest little clinical heterogeneity. This is because the curated data on monogenic diseases tend to be structurally simpler than their multigenic counterparts. The entries of many attributes such as gene symbol, variant type, ethnicity, and clinical phenotype tend to be more homogenous in monogenic diseases (*see* **Note 4**). For multigenic diseases, flexibility in data querying is uppermost because of the heterogeneity of the underlying data and this is where the strength of fine-grain mapping lies.

2. Programming skill available

   Coarse grain database can be queried with simpler Structured Query Language (SQL) commands since the curated data are likely to be stored in fewer tables and fewer attributes than its fine grain counterpart. A fine grain database can be a collection of more than 10 tables and over 50 attributes. Hence, it will demand a higher level of proficiency in SQL commands.

2.3.1.1. Tutorials on Developing a DNA Variant Database

In the following tutorials, the reader will learn step-by-step the process of designing a database schema on paper before implementing it on a DBMS, and then populate the resulting database with some sample data. The aim is to give the reader an idea on how to build a database using one of the most popular open-source DBMS, MySQL, on Windows® XP. Designing a web interface for user query and programming the web application required for making the database web accessible are beyond the scope of this tutorial.

2.3.1.1.1. Tutorial 1: Mapping Domain Knowledge to Class-Responsibility-Collaboration (CRC) cards. In this section, a hypothetical case will be used to illustrate the process of mapping domain knowledge on a locus specific disease to a series of CRC cards on paper. CRC cards have previously been used for modeling the property, e.g., protein name, gene symbol, cytogenetic location, etc., and functionality, e.g., known pathways, type of biochemical reactions, etc., of an individual protein, and its interactions with other proteins ***(15)***. Here, the purpose of CRC cards is to model the properties of each class, its functionality in the database, and its relationship with other classes.

1. Determine what the classes are and what will be the relationship between classes. For example, the statement "Gene X carries DNA variants" describes a relationship between the classes *Locus* and *DNA variant*. Generally speaking, the core and auxiliary data mentioned in **section 2.2.2.1** can be mapped to the classes *Locus*, *DNA variant*, *Citation*, *Population*, and *Kindred*.
2. Determine what are the attributes of each class. The attributes should describe the property of the class. Each datum should map to an attribute unique to a particular class.
3. The CRC card for the class *Locus* is as follows:

   Class: *Locus*

   Attributes: OMIM identifier, GenBank identifier, gene symbol, gene name, cytogenetic position, locus type (*see* **Note 5**)

   Responsibility: This class stores the identifiers of a genetic locus. A locus may carry more than one DNA variant (*see* **Note 6**).

   Collaboration: *DNA variant*

4. The CRC card for the class *DNA variant* is as follows:
   Class: *DNA variant*
   Attributes: intragenic position, gDNA change, cDNA change, residue change, variant type (*see* **Note 7**)
   Responsibility: This class stores the position and the type of nucleotide changes on the gDNA and cDNA levels and the corresponding residue change. The data in this class are annotated in HUGO nomenclatures ***(4,5)***.
   Collaboration: *Citation*, *Locus*, *Population*
5. The CRC card for the class *Population* is as follows:
   Class: *Population*
   Attributes: continent, nation, ethnicity (*see* **Note 8**)
   Responsibility: This class stores data on the geographical location and the ethnicity of the human populations known to carry a set of DNA variants. A population may carry more than one DNA variant and the reverse may also be true (*see* **Note 9**).
   Collaboration: *DNA variant*, *Kindred*
6. The CRC card for the class *Kindred* is as follows:
   Class: *Kindred*
   Attributes: phenotype, number of afflicted subjects, number of asymptomatic carriers
   Responsibility: This class stores the kindred data associated with a DNA variant (*see* **Note 10**). A DNA variant may be found in one or more kindreds and a kindred may carry more than one variant.
   Collaboration: *Population*
7. The CRC card for the class *Citation* is as follows:
   Class: *Citation*
   Attributes: PubMed ID, author, publication
   Responsibility: This class stores the bibliographic data of a DNA variant. A DNA variant carried by several populations is usually reported in one or more publications. If the DNA variant is associated with multiple kindreds, both are usually reported together in the same publication (*see* **Note 11**).
   Collaboration: *DNA Variant*

2.3.1.2. Implementing the Database Schema

The procedure introduced in tutorial 2 can be used with both open-source, e.g., PostgreSQL, and commercial DBMSes e.g. Oracle® and Microsoft SQL Server®, other than MySQL with some modifications on step 4. It should be noted that the syntax of Data Definition Language (DDL) often differs slightly from one DBMS to another. It is recommended that the user should check the reference manual of other DBMSes for details. The reference manual for MySQL DBMS can be found in ***(16)***.

2.3.1.2.1. Tutorial 2: Mapping the CRC Cards to a Database Schema. Using the CRC cards built in the last section, one can build the database schema in DDL as follows:

1. Add primary key to each class. This primary key will become the identifier for each record of a particular class. The primary key and its sample entry for each class are shown in **Table 1**.
2. Determine the data type of every attribute in each class **(Table 2)**.
3. Convert CRC card into a conceptual schema of a sample database named Locus-Variants as shown in **Fig. 2** (*see* **Note 12**).
4. Convert the conceptual schema to a MySQL database schema in DDL clauses as shown in **Fig. 3** (*see* **Note 13**). Each class serves as a template for building a table. All database-reserved keywords are in uppercase.
5. Save the above DDL clauses in a text file named locusvariants.sql. (*see* **Note 14**)

2.3.1.2.2. Tutorial 3: Implement Database Schema in MySQL DBMS. After building the database schema in Data Definition Language, the next step will be to implement the schema in MySQL DBMS. The procedure is as follows:

1. Open the MySQL Query Browser under the MySQL menu **(Fig. 4)**
2. Login as **root** and enter the appropriate password. Leave the Default Schema box in the Login dialog blank **(Fig. 5)** (*see* **Note 15**).
3. Load the schema into the database by choosing the Open Script item under the File menu **(Fig. 6)**.
4. Then choose the locusvariants.sql from the stored directory path in Windows. For example, if locusvariants.sql is being stored under the directory My Documents, then the directory path should be C:/my documents.
5. Execute the script by clicking on the button Execute once in the Menu bar **(Fig. 7a)**. The reader should immediately see the database locusvariants and the various tables appearing in the Schemata sub-window **(Fig. 7b)**.

Table 1
Primary classes and primary keys of each class

Classes	Primary Keys	Data Types
Locus	mim_id	long integer, e.g., 144241
DNA_variant	dno	Text of 7 characters (alphanumerics), e.g., DVR0001
Population	pno	Text of 4 characters (alphanumerics), e.g., P001
Kindred	cno	Text of 3 characters (alphanumerics), e.g., C01
Citation	pubmed_id	long integer, e.g., 562890

Table 2
Attributes in each class and their definitions

Classes	Attributes	Definitions	Data Types
Locus	gb_id	GenBank identifier for the genomic sequence of the locus.	Text of up to 15 characters, e.g., NM_100034
Locus	gene_symbol	HUGO-assigned gene symbol of the locus.	Text of up to 10 characters, e.g., CDK4
Locus	genename	The full name of the gene.	Text of up to 50 characters e.g. cyclin-dependent kinase 4
Locus	chr_no	The chromosome at which the locus is.	Text of up to 2 characters e.g. 11
Locus	cytoband	The chromosomal location of the locus.	Text of up to 12 characters e.g. 12q14.1
Locus	locus_type	The classification of the locus according to the type of gene product expressed.	Text of up to 15 characters e.g. protein-coding
DNA_variant	gene_element	The region within the open reading frame where the DNA variant is being located.	Text of up to 8 characters, e.g., exon 10
DNA_variant	gdna_change	A nucleotide change in the genomic DNA in reference to the GenBank sequence.	Text of up to 20 characters, e.g., g. 247C>T
DNA_variant	cdna change	A nucleotide change in the complementary DNA as a result of the gDNA change	Text of up to 40 characters, e.g., c.143C>T
DNA variant	residue_change	An residual change in the polypeptide as a result of the cDNA change.	Text of up to 20 characters, e.g., p.P48L
DNA variant	variant_type	The type of gDNA change concerned	Text of up to 30 characters, e.g., point transition
Population	continent	The name of the continent where the gDNA variant was discovered	Text of up to 15 characters, e.g., Australasia

Population	nation	The name of the nation or country where the gDNA variant was discovered	Text of up to 15 characters, e.g., New Zealand
Population	ethnicity	The anthropological classification of the affected population	Text of up to 30 characters, e.g., Caucasoid, Anglo-Celtic
Kindred	phenotype	The clinical condition observed in the affected population	Text of up to 40 characters, e.g., pancreatic carcinoma
Kindred	nr_afflicted	The number of subjects expressing the clinical phenotype	short integer, e.g., 12
Kindred	nr_carriers	The number of asymptomatic carriers	short integer, e.g., 12
Citation	pubmed_id	The PubMed identifier of the publication	long integer, e.g., 530342
Citation	authors	The authors' names on the publication	Text of up to 100 characters, e.g., Smith J
Citation	publication	The journal title, year, volume No, page No.	Text of up to 50 characters, e.g., Am J Hum Genet 2004, **100:**90–95

2.3.1.2.3. Tutorial 4: Populating the Database with Curated Data. Now that the database has been created, the next step is to populate it with data. The easiest approach is to store input data for each table as a tab-delimited text file and then use SQL commands to load data into the appropriate table.

1. Create nine text files using Wordpad in Windows and save each as locus.txt, variant.txt, population.txt, kindred.txt, citation.txt, loc2var.txt, var2popl.txt, popl2kin.txt, and var2cit.txt, respectively.
2. Enter sample data shown in **Fig. 8** to an appropriate text file and save it. In each file, the number of columns should match the exact number of attributes of a particular table. For example, the file locus.txt contains seven tab-delimited columns that map to the attributes mim_id, gb_id, gene_symbol, genename, chr_no, cytoband, and locus_type. There should be a tab space between every pair of data in the same row.
3. Create another text file using Wordpad in Windows and save it as loaddata.sql.
4. Type in a list of commands that will automatically load data from the appropriate text file to the appropriate table. Each command has the following syntax:

```
Database LocusVariants {
        Class Locus {
                Attribute: mim_id            long integer     primary key
                Attribute: gb_id             text
                Attribute: gene_symbol       text
                Attribute: genename          text
                Attribute: chr_no            text
                Attribute: cytoband          text
                Attribute: locus_type        text
        }
        Class DNA_variant {
                Attribute: dno               text             primary key
                Attribute: gene_element      text
                Attribute: gdna_change       text
                Attribute: cdna_change       text
                Attribute: residue_change    text
                Attribute: variant_type      text
        }
        Class Population {
                Attribute: pno               text             primary key
                Attribute: continent         text
                Attribute: nation            text
                Attribute: ethnicity         text
        }
        Class Kindred {
                Attribute: cno               text             primary key
                Attribute: phenotype         text
                Attribute: nr_afflicted      short integer
                Attribute: nr_carriers       short integer
        }
        Class Citation {
                Attribute: pubmed_id         long integer     primary key
                Attribute: authors           text
                Attribute: publication       text
        }
        Class Locus_2_Variant {
                Attribute: mim_ref foreign key→ Locus.mim_id
                Attribute: dno_ref foreign key→ DNA_variant.dno
        }
        Class Variant_2_Population {
                Attribute: dno_ref foreign key→ DNA_variant.dno
                Attribute: pno_ref foreign key→ Population.pno
        }
        Class Population_2_Kindred {
                Attribute: pno_ref foreign key→ population.pno
                Attribute: cno_ref foreign key→ kindred.cno
        }
        Class Variant_2_Citation {
                Attribute: dno_ref foreign key→ DNA_variant.dno
                Attribute: pubmed_ref foreign key→ Citation.pubmed_id
        }
}
```

Fig. 2. The conceptual schema for the database LocusVariants.

```
CREATE DATABASE IF NOT EXISTS locusvariants;
USE locusvariants;
CREATE TABLE locus (
        mim_id INT(8) PRIMARY KEY,
        gb_id VARCHAR(15) NOT NULL,
        gene_symbol VARCHAR(10),
        genename VARCHAR(50),
        chr_no VARCHAR(2),
        cytoband VARCHAR(12),
        locus_type VARCHAR(15),
        UNIQUE(mim_id, gene_symbol)
);
CREATE TABLE dna_variant (
        dno VARCHAR(7) PRIMARY KEY,
        gene_element VARCHAR(8),
        gdna_change VARCHAR(20) NOT NULL,
        cdna_change VARCHAR(40) NOT NULL,
        residue_change VARCHAR(20),
        variant_type VARCHAR(30) NOT NULL,
        UNIQUE(dno)
);
CREATE TABLE population (
        pno VARCHAR(4) PRIMARY KEY,
        continent VARCHAR(15),
        nation VARCHAR(15),
        ethnicity VARCHAR(30),
        UNIQUE(pno)
);
CREATE TABLE kindred (
        cno VARCHAR(3) PRIMARY KEY,
        phenotype VARCHAR(40),
        nr_afflicted INT(2),
        nr_carriers INT(2),
        UNIQUE(cno)
);
CREATE TABLE citation (
        pubmed_id INT(8) PRIMARY KEY,
        authors VARCHAR(100) NOT NULL,
        publication VARCHAR(50) NOT NULL,
        UNIQUE (pubmed_id)
);
CREATE TABLE locus2variant (
        mim_ref INT(8) NOT NULL,
        dno_ref VARCHAR(7) NOT NULL,
        FOREIGN KEY (mim_ref) REFERENCES locus(mim_id),
        FOREIGN KEY (dno_ref) REFERENCES dna_variant(dno)
);
```

Fig. 3. (*Continued*)

```
CREATE TABLE variant2population (
        dno_ref VARCHAR(7) NOT NULL,
        pno_ref VARCHAR(4) NOT NULL,
        FOREIGN KEY (dno_ref) REFERENCES dna_variant(dno),
        FOREIGN KEY (pno_ref) REFERENCES population(pno)
);
CREATE TABLE population2kindred (
        pno_ref VARCHAR(4) NOT NULL,
        cno_ref VARCHAR(3) NOT NULL,
        FOREIGN KEY(pno_ref) REFERENCES population(pno),
        FOREIGN KEY(cno_ref) REFERENCES kindred(cno)
);
CREATE TABLE variant2citation (
        dno_ref VARCHAR(7) NOT NULL,
        pubmed_ref INT(8) NOT NULL,
        FOREIGN KEY(dno_ref) REFERENCES dna_variant(dno),
        FOREIGN KEY(pubmed_ref) REFERENCES citation(pubmed_id)
);
```

Fig. 3. The DDL schema for the database LocusVariants.

   LOAD DATA INFILE 'c:\\path\\<*filename.txt*>' INTO TABLE <*table_name*> LINES TERMINATED BY '\n';
   The path is any directory path in Windows, the <*filename.txt*> is the name of the text file, and the <*table_name*> is the name of the target table. For example, if locus.txt is being stored in the directory My Documents and its target table is locus, the above command line will be
   LOAD DATA INFILE 'c:\\my documents\\locus.txt' INTO TABLE LOCUS LINES TERMINATED BY '\n';

5. Using **Table 3** as a guide, repeat step 4 for each of the text files (*see* **Note 16**). Eventually, the loaddata.sql file should contain a list of commands like:
   USE locusvariants;
   LOAD DATA INFILE 'c:\\my documents\\locus.txt' INTO TABLE locus LINES TERMINATED BY '\n';
   LOAD DATA INFILE 'c:\\my documents\\variant.txt' INTO TABLE dna_variant LINES TERMINATED BY '\n';
   LOAD DATA INFILE 'c:\\my documents\\kindred.txt' INTO TABLE KINDRED LINES TERMINATED BY '\n';
6. Load the file loaddata.sql into the database by repeating steps 3–5 in tutorial 3.

2.3.1.2.4. Tutorial 5: Searching Data from the Database. With the database populated, one has to know how to search the database and piece the extracted data together in order to give an informative view about the variants of a locus. This tutorial will equip the readers with some basic skills in formulating SQL queries.

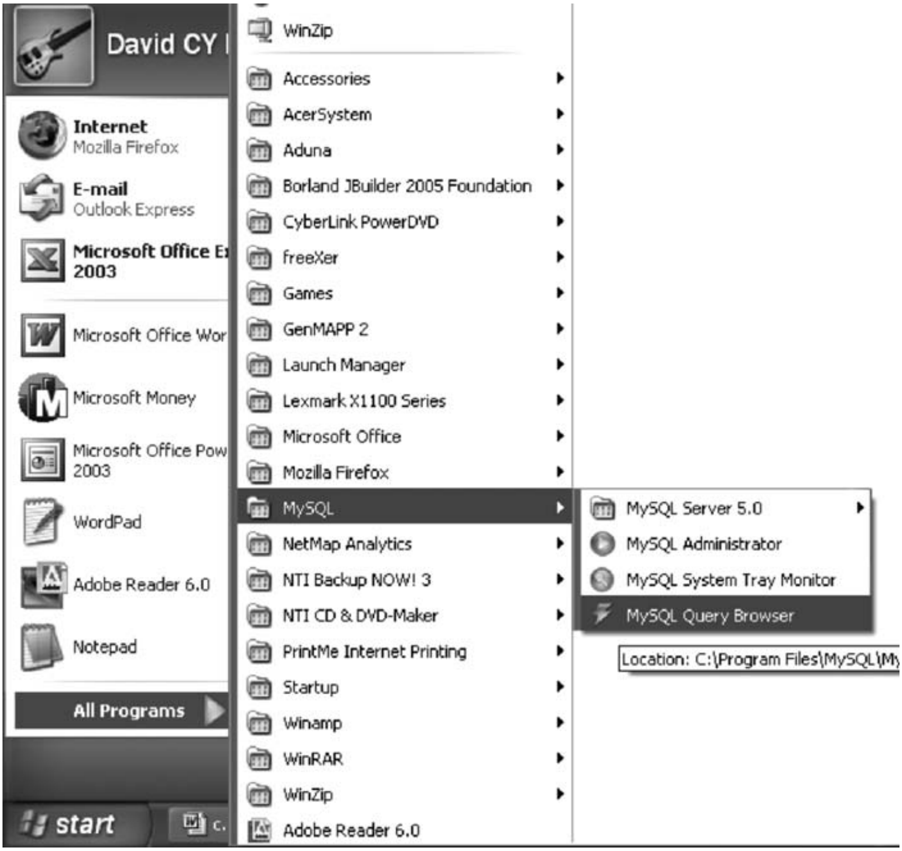
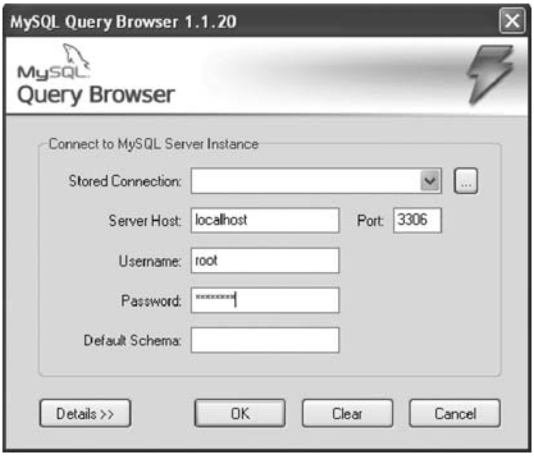


Fig. 4. Screenshot of MySQL menu in Windows® XP.

The formulation of an SQL query involves three basic steps: (1) determine the list of tables that will be joined together, (2) determine the list of attributes to be queried from the tables, and (3) determine the attributes that will serve as the search criteria (optional). In this tutorial, the author will use two examples as an illustration.

Example 1: Find all the gDNA variants carried in the *CDKN2A* gene.

1. The tables needed for this example are locus, locus2variant, and dna_variant.
2. The attributes needed are mim_id, gene_symbol, genename, cytoband, locus_type, gene_element, gdna_change, and variant_type. There is no need to define gene_symbol as a search criterion because *CDKN2A* is the only entry stored as gene_symbol in the locus table.



Stored Connection:

Server Host: localhost Port: 3306

Username: root

Password: ********

Default Schema:

Fig. 5. Screenshot of MySQL login dialog.

3. To perform the required database search, the SQL command SELECT…FROM is required. This command has the following syntax:
   ```
   SELECT <list of attributes>
   FROM <list of tables>
   WHERE <attribute1 = attribute2 OR attribute1 = 'value1'>
   AND <attribute1 = attribute2 OR attribute2 = 'value2'>
   ```
   Thus the appropriate SELECT…FROM command should be
   ```
   SELECT l.mim_id, l.gene_symbol, l.genename, l.cytoband, l.locus_type, v.dno,
   v.gene_element, v.gdna_change, v.cdna_change, v.residue_change, v.variant_type
   FROM locus l, locus2variant lv, dna_variant v
   WHERE l.mim_id = lv.mim_ref
   AND v.dno = lv.dno_ref;
   ```
4. Type the above command into the box next to the Execute button in the Menu bar **(Fig. 9)**.
5. Click on the Execute button to execute the command and the reader should see the result in the Resultset sub-window **(Fig. 9)**.

Example 2: Find the auxiliary data related to all the gDNA variants found in the country Australia.

1. The tables needed for this example are dna_variant, population, kindred, citation, variant2population, population2kindred, and variant2citation.
2. The attributes needed are dno, gene_element, gdna_change, continent, nation, ethnicity, cno, nr_afflicted, nr_carriers, phenotype, pubmed_id and publication.
3. The search criterion is the attribute nation that has an entry equal to Australia.

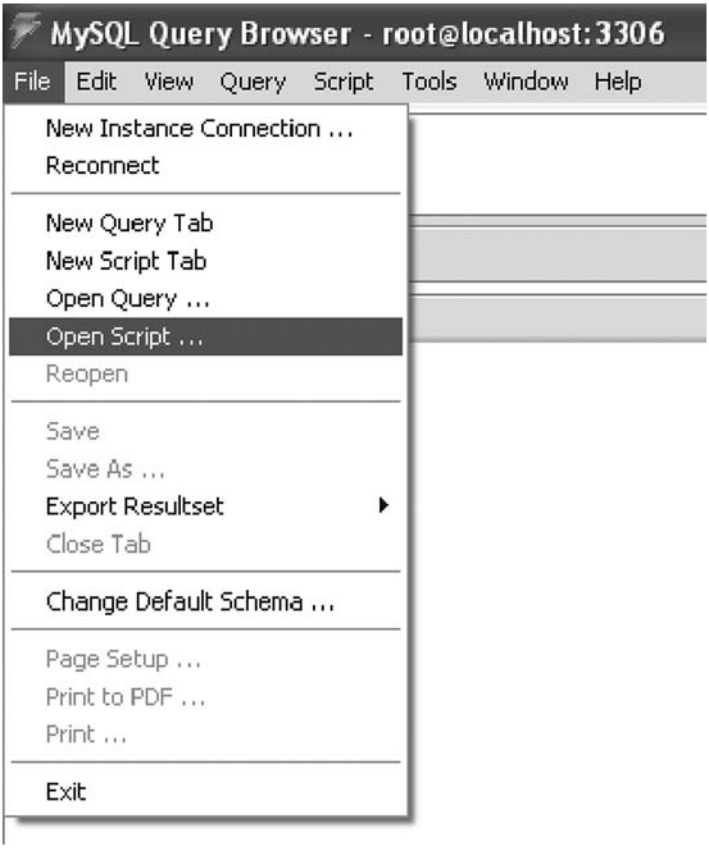
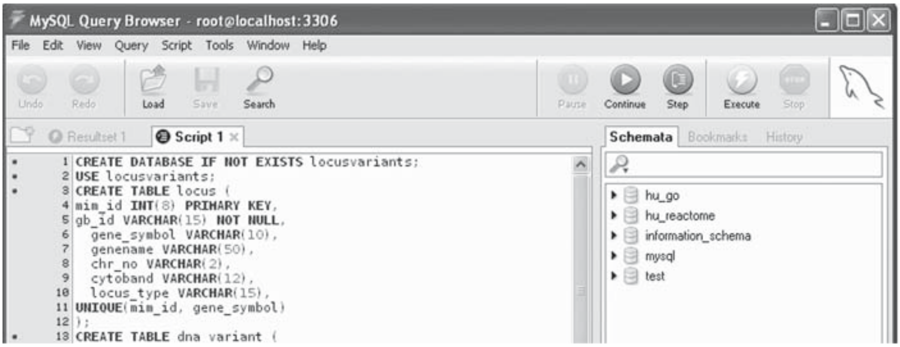


Fig. 6. Screenshot of the File menu in MySQL Query Browser window.

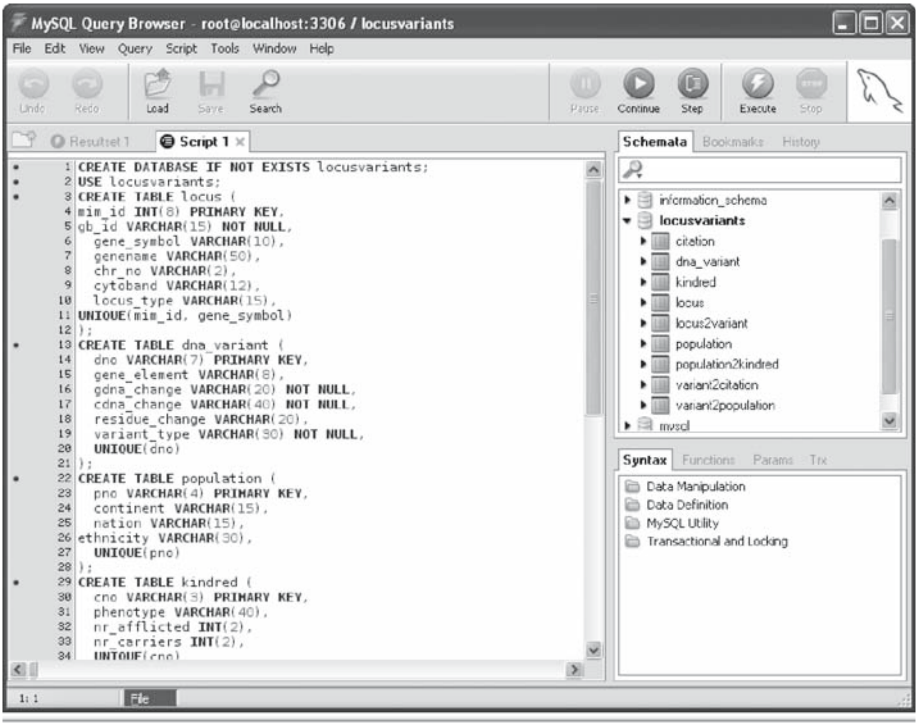
4. The SQL command should therefore be: SELECT d.dno, d.gene_element, d.gdna_change, p.continent, p.nation, p.ethnicity,k.cno,k.nr_afflicted,k.nr_carriers, k.phenotype,c.pubmed_id, c.publication
   FROM dna_variant d,population p,kindred k,citation c,variant2population vp,variant2citation vc, population2kindred pk
   WHERE p.nation = ‘Australia’
   AND vp.dno_ref = d.dno
   AND vp.pno_ref = p.pno
   AND vc.dno_ref = d.dno
   AND vc.pubmed_ref = c.pubmed_id
   AND pk.pno_ref = p.pno
   AND pk.cno_ref = k.cno;

(a)


MySQL Query Browser - root@localhost:3306
File Edit View Query Script Tools Window Help
Undo Redo Load Save Search Pause Continue Step Execute Stop
Resultset 1
Script 1
Schemata Bookmarks History
hu_go
hu_reactome
information_schema
mysql
test

```
CREATE DATABASE IF NOT EXISTS locusvariants;
USE locusvariants;
CREATE TABLE locus (
mim_id INT(8) PRIMARY KEY,
gb_id VARCHAR(15) NOT NULL,
  gene_symbol VARCHAR(10),
  genename VARCHAR(50),
  chr_no VARCHAR(2),
  cytoband VARCHAR(12),
  locus_type VARCHAR(15),
UNIQUE(mim_id, gene_symbol)
);
CREATE TABLE dna_variant (
```

(b)


MySQL Query Browser - root@localhost:3306 / locusvariants
File Edit View Query Script Tools Window Help
Undo Redo Load Save Search Pause Continue Step Execute Stop
Resultset 1
Script 1
Schemata Bookmarks History
information_schema
locusvariants
citation
dna_variant
kindred
locus
locus2variant
population
population2kindred
variant2citation
variant2population
mysql
Syntax Functions Params Trx
Data Manipulation
Data Definition
MySQL Utility
Transactional and Locking
1: 1
File

```
CREATE DATABASE IF NOT EXISTS locusvariants;
USE locusvariants;
CREATE TABLE locus (
mim_id INT(8) PRIMARY KEY,
gb_id VARCHAR(15) NOT NULL,
  gene_symbol VARCHAR(10),
  genename VARCHAR(50),
  chr_no VARCHAR(2),
  cytoband VARCHAR(12),
  locus_type VARCHAR(15),
UNIQUE(mim_id, gene_symbol)
);
CREATE TABLE dna_variant (
  dno VARCHAR(7) PRIMARY KEY,
  gene_element VARCHAR(8),
  gdna_change VARCHAR(20) NOT NULL,
  cdna_change VARCHAR(40) NOT NULL,
  residue_change VARCHAR(20),
  variant_type VARCHAR(30) NOT NULL,
  UNIQUE(dno)
);
CREATE TABLE population (
  pno VARCHAR(4) PRIMARY KEY,
  continent VARCHAR(15),
  nation VARCHAR(15),
ethnicity VARCHAR(30),
  UNIQUE(pno)
);
CREATE TABLE kindred (
  cno VARCHAR(3) PRIMARY KEY,
  phenotype VARCHAR(40),
  nr_afflicted INT(2),
  nr_carriers INT(2),
  UNIQUE(cno)
```

Fig. 7. Screenshot **(a)** before executing and **(b)** after executing the locusvariants.txt file.

(a)	600160	AF527803	CDKN2A	cyclin-dependent inhibitor 2A	kinase	9	9p21	protein-coding

(b)	DVR0001	exon 1	g.1247C>T	c.143C>T	p.P48L	point transition
	DVR0002	exon 1	g.1253A>G	c.149A>G	p.Q50R	point transition
	DVR0003	exon 2	g.1269-1289del	c.167-197del	p.G67fsX145	indel deletion

(c)	P001	Australasia	Australia	Caucasoid Anglo-Celtic
	P002	Europe	Italy	Caucasoid Italian
	P003	Europe	France	Caucasoid French

(d)	C01	pancreatic carcinoma	20	0
	C02	cutaneous malignant melanoma	16	1
	C03	cutaneous malignant melanoma	39	4
	C04	cutaneous malignant melanoma	7	1

(e)	93521	Henderson L, Waters J, Lanting F, Meydon DE	Intl J Cancer Genet 1994. 100:1002–1004
	405685	Frizzi C, Carlos M	Oncogene 2003. 20:917–920
	562890	Pierre J, Fromm P, Ascome M	J Cancer Res 2006. 120:10002–10006

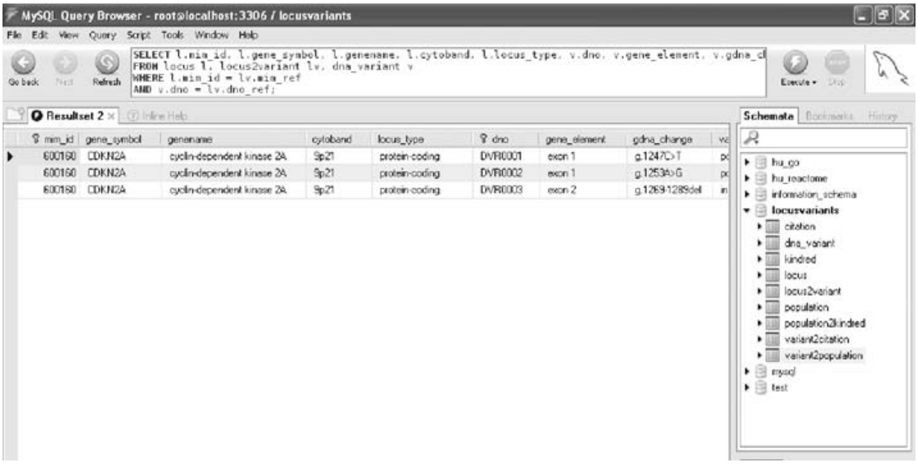
(f)	600160	DVR0001
	600160	DVR0002
	600160	DVR0003

(g)	DVR0001	P002
	DVR0001	P003
	DVR0002	P001
	DVR0003	P001

(h)	P001	C01
	P001	C02
	P002	C03
	P003	C04

(i)	DVR0001	405685
	DVR0001	562890
	DVR0002	93521
	DVR0003	93521

Fig. 8. Sample data input for the nine tab-delimited files: (a) locus.txt, (b) variant.txt, (c) loc2var.txt, (d) population.txt, (e) kindred.txt, (f) var2popl.txt, (g) popl2kin.txt, (h) citation.txt, and (i) var2cit.txt. Each text file corresponds to one of the nine tables in the database **(Table 3)**.



```
SELECT l.mim_id, l.gene_symbol, l.genename, l.cytoband, l.locus_type, v.dno, v.gene_element, v.gdna_cl
FROM locus l, locus2variant lv, dna_variant v
WHERE l.mim_id = lv.mim_ref
AND v.dno = lv.dno_ref;
```

mim_id	gene_symbol	genename	cytoband	locus_type	dno	gene_element	gdna_change	va
600160	CDKN2A	cyclin-dependent kinase 2A	9p21	protein-coding	DVR0001	exon 1	g.1247C>T	pc
600160	CDKN2A	cyclin-dependent kinase 2A	9p21	protein-coding	DVR0002	exon 1	g.1253A>G	pc
600160	CDKN2A	cyclin-dependent kinase 2A	9p21	protein-coding	DVR0003	exon 2	g.1269-1289del	in

Fig. 9. Screenshot after executing the SQL query for example 1.

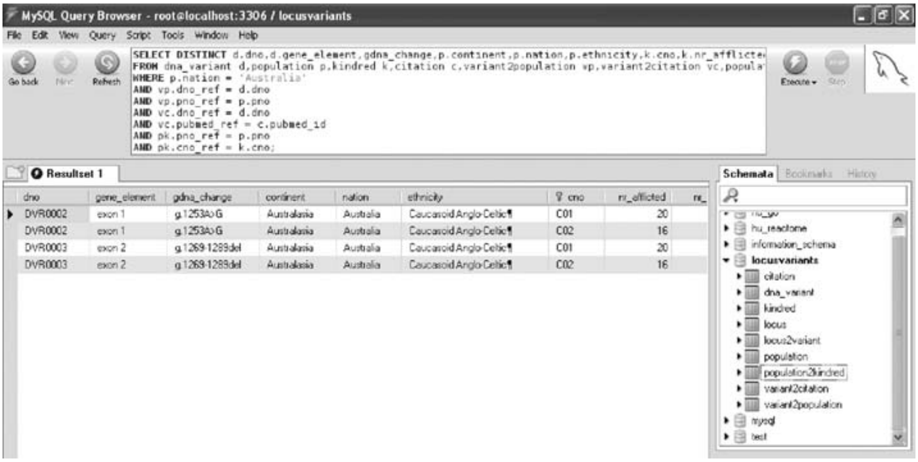
5. Type the above command into the box next to the Execute button in the Menu bar **(Fig. 10)**.
6. Click on the Execute button to execute the command and the reader should see the result in the Resultset sub-window **(Fig. 10)**.

2.3.1.3. Hosting the Database Online

To make the database web accessible, it has to be connected to a web server via a web application. Examples of commonly used open-source web servers

Table 3
Schema for mapping the text files to the database tables

Text Files	Tables
locus.txt	locus
variant.txt	dna_variant
population.txt	population
kindred.txt	kindred
citation.txt	citation
loc2var.txt	locus2variant
var2popl.txt	variant2population
popl2kin.txt	population2kindred
var2cit.txt	variant2citation



```
SELECT DISTINCT d.dno,d.gene_element,gdna_change,p.continent,p.nation,p.ethnicity,k.cno,k.nr_afflicte
FROM dna_variant d,population p,kindred k,citation c,variant2population vp,variant2citation vc,popula
WHERE p.nation = 'Australia'
AND vp.dno_ref = d.dno
AND vp.pno_ref = p.pno
AND vc.dno_ref = d.dno
AND vc.pubmed_ref = c.pubmed_id
AND pk.pno_ref = p.pno
AND pk.cno_ref = k.cno;
```

dno	gene_element	gdna_change	continent	nation	ethnicity	cno	nr_afflicted	nr_
DVR0002	exon 1	g.1253A>G	Australasia	Australia	Caucasoid Anglo-Celtic¶	C01	20	
DVR0002	exon 1	g.1253A>G	Australasia	Australia	Caucasoid Anglo-Celtic¶	C02	16	
DVR0003	exon 2	g.1269-1289del	Australasia	Australia	Caucasoid Anglo-Celtic¶	C01	20	
DVR0003	exon 2	g.1269-1289del	Australasia	Australia	Caucasoid Anglo-Celtic¶	C02	16	

Fig. 10. Screenshot after executing the SQL query for example 2.

include JBoss *(17)*, Apache *(18)* and Tomcat-Jakarta *(19)*. Although it is beyond the scope of this tutorial to teach Web programming, it will be helpful for the reader to know what the two most common types of web application are, so that one can appreciate the programming skill required.

The most established implementation of web application is the common gateway interface (CGI). A CGI application is a module that receives the user's request from the Web server (*see* **Note 17**). The CGI then forwards SQL queries to the database and sends the result back to the server, typically in Hypertext Text Markup Language (HTML). In turn the server sends the HTML back to the user's browser which displays the HTML as a Web page on the fly.

Although CGI can be written in any language, most have been written in Perl. It is an interpreted language designed for processing text data. As such, the Web server has to execute the Perl interpreter and reload the CGI application into memory for handling every user's request. For this reason, CGI is a single thread single process method.

A more efficient alternative to CGI is server-side Java. In place of the CGI, a Java application receives the user's request from the Web server and forwards SQL queries to the database, but that is where their similarity ends. Java is a compiled language. Any Java application has to be compiled into the bytecode form to make it executable by the Java Virtual Machine (JVM). Once compiled, the application is loaded into memory upon the first user's request. It persists in memory, handling more users' requests, until it becomes idle for a limited time span. Then it will be deleted from memory by the JVM, a process known as garbage collection. This multi-thread single process is much more efficient than its CGI counterpart ***(20)***. As to which hosting method should be used, the decision depends largely on the (1) complexity of the data, as discussed in **section 2.3.1**, and (2) functionalities of the web interface, e.g., the provision of data filtering by multiple criteria and the provision of web interface for data curation. An example for each implementation can be found in ***(21,22)***.

2.4. Transition

2.4.1. Deploying the DNA Variant Database

One method of deploying a web accessible database is the two-tier master/slave configuration. This requires two computers that have been installed with the same DBMS and web server. The internal tier located behind the firewall has been installed with the master copy of the database and is accessible only to the curator. The tier outside the firewall has been installed with the slave copy of the database which is accessible to the public. The firewall that separates the two tiers is configured to allow data streaming from the internal tier to the public tier but not vice versa. If the slave copy in the public tier is being corrupted, it will be replaced with another copy duplicated from the master. Hence, the biggest advantage of the master/slave configuration is data security. However, the CPU power and memory capacity in each tier has to be shared between the DBMS and the Web server on board. This is a disadvantage to the public tier since it may have to handle increasing volume of users' requests during its service lifetime. In other words, the configuration lacks scalability.

A modification of the above method of deployment is the three-tier master/slave configuration in which each tier will specialize in providing a single service. The slave and the master copy of the database will be running on two internal tiers respectively but the Web server will be running on the public tier. This configuration resolves the limitation of its two-tier counterpart, but at a higher cost of maintenance since it now requires three computers instead of two.

2.4.2. Promoting the Web Accessible Variant Database to the User Community

It is highly recommended that the HGVS be notified on any newly hosted variant database so that the organization can add the web address of the new database to its catalogue. It will also be useful to contact the administrators of HGMD® to see if they are interested in creating hyperlinks to one's variant database. Finally, one should publish the course and results of the project in a peer-reviewed journal as a form of recognition.

3. Notes

1. In this chapter, the term "DNA variant" stands for any intragenic nucleotide variant.
2. A web application is a collection of programs that connects the database to the Web server. *See* **section 2.3.1.3** for further elaboration.
3. A class is a data structure designed for modeling a real world object such as a locus, a gDNA variant, or a population. The basic element of a class is an attribute and the property of a class is represented by a collection of attributes.
4. While this statement is generally true, certain monogenic diseases can express very broad clinical heterogeneity while genetically very homogenous. Cystic fibrosis is one example with over two-thirds of the affected being homozygous carriers of F508del in the *CFTR* gene but with no fewer than six clinical phenotypes being observed. It has recently been suggested that TNFα receptor and certain sodium channel genes could be modifiers ***(23)***. If the reader wants to include variants of modifier loci to a monogenic disease-centred database, the data may eventually exhibit the heterogeneity seen with multigenic disease data and coarse-grain mapping may not be the best approach.
5. Although a majority of loci stored in published variant databases are protein-coding, a minority could be RNA-coding.
6. Since a single locus can carry multiple DNA variants, the cardinality is 1:n. The same does not necessarily hold in the reverse which is usually in a cardinality of 1:1 because each DNA variant should map to one distinct locus.
7. In the case of RNAi and small RNA genes, the attribute residue change can be omitted. The entries for the variant type can be found in the EBI mutation event keyword schema ***(24)***.
8. The classification of ethnicity can be found in the EBI Mutation Database ***(24)***. Recommendations and the naming of countries can be found in the US Central Intelligence Agency's World Factbook ***(25)***.
9. The cardinality of the relationship *Population*-to-*DNA variant* is 1:n. The reverse would also be true. This type of relationship is known as bidirectional.
10. One would argue that *Kindred* should be merged with *Population*. Whether it is necessary to store data relating to kindred in a separate table depends on the complexity of the disease. If the *Population*-to-*Kindred* is a 1:1 relationship, the two tables can be merged into one.
11. The more frequently a particular DNA variant is being published, the more likely that it is a mutation hotspot. Furthermore, multiple reporting of a particular DNA variant inevitably increases its authenticity.
12. The problem with mapping 1:n relationships to the tabular format of a relational database schema is the generation of data redundancy. For example, if Locus A carries six DNA variants, this will mean that the only variable in the *Locus* class will be the foreign key dno_ref that maps to the primary key dno of the *DNA variant* class and the same entry for the rest of the attributes in *Locus* will be stored repeatedly for six times. The solution is to store the primary keys of the two

classes in a separate class called *Locus_2_DNA_variant*. Tables resulting from classes of this type are known as intermediate tables **(Fig. 2)**.

13. Each CREATE TABLE clause must end with a semi-colon.
14. All the text files created in these tutorials should be saved as the file type Text Document, which is a plain ASCII format readable by MySQL DBMS.
15. If the MySQL DBMS has been installed on the reader's own computer, the password for the root account should have already been decided during the installation process.
16. It is important to load the data tables, e.g., dna_variant, population, and etc., into the database before the intermediate tables e.g. variant2population, in order to avoid the following type of errors: Cannot add or update a child row: a foreign key constraint fails ('locusvariants/variant2population', CONSTRAINT 'variant2population_ibfk_2' FOREIGN KEY('pno_ref') REFERENCES 'population' ('pno')) In the example above, the MySQL database engine failed to match the foreign key pno_ref in the intermediate table variant2population to the primary key in the population table simply because the latter has yet to be populated.
17. An application is a collection of programs with each typically performing a single function.

References

1. Human Gene Mutation Database. (2006) Url: http://www.hgmd.cf.ac.uk/.
2. Human Genome Variation Society. (2006) Url: http://www.hgvs.org/.
3. Shaw, D. (2001) Extreme programming in context. *Syst. Developer* **2**, 20–26.
4. Antonarakis, S. E., and Nomenclature Working Group. (1998) Recommendations for a nomenclature system for human gene mutations. *Hum. Mutat.* **11**, 1–3.
5. den Dunnen, J. T., and Antonarakis, S. E. (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **15**, 7–12.
6. Graves, M. (2002) *Designing XML Databases*. Prentice Hall, Upper Saddle River, NJ, pp. 163–165.
7. Alders, M., Jongbloed, R., Deelan, W., van den Wijngaard, A., Doevendans, P., Ten Cate, F., et al. (2003) The 2373insG mutation in the MYBPC3 gene is a founder mutation, which accounts for nearly one-fourth of the HCM cases in the Netherlands. *Eur. Heart J.* **24**, 1848–1853.
8. Hougs, L., Havndrup, O., Bundgaard, H., Kober, L., Vuust, J., Larsen, L. A., et al. (2005) One third of Danish hypertrophic cardiomyopathy patients with MYH7 mutations have mutations in MYH7 rod region. *Eur. J. Hum. Genet.* **13**, 161–165. *Eur. J. Hum. Genet. editorial. ibid.* **13**, 694.
9. van der Velden, P. A., Sandkuijl, L. A., Bergman, W., Pavel, S., van Mourik, L., Frants, R. R., et al. (2001) Melanocortin receptor 1 variant R151C modifies the melanoma risk in Dutch families with melanoma. *Am. J. Hum. Genet.* **69**, 774–779.

10. Jen, J., Yue, Q., Nelson, S. F., Yu, H., Litt, M., Nutt, J., et al. (1999) A novel nonsense mutation in CACNA1A causes episodic ataxia and hemiplegia. *Neurology* **53**, 34–37.
11. Locke III, G. R. , Ackerman, M. J., Zinsmeister, A. R., Thapa, P., and Farrugia, G. (2006) Gastrointestinal symptoms in families of patients with an SCN5A-encoded cardiac channelopathy: evidence of an intestinal channelopathy. *Am. J. Gasteroenterol.* **101**, 1299–1304.
12. Waldmuller, S., Sakthivel, S., Saadi, A. V., Selignow, C., Rakesh, P. G., Golubenko, M., et al. (2003) Novel deletions in MYH7 and MYBPC3 identified in Indian families with familial hypertrophic cardiomyopathy. *J. Mol. Cell. Cardiol.* **35**, 623–636.
13. Huang, X., Song, L., Ma, A. Q., Gao, J., Zheng, W., Wu, C. W., et al. (2001) A malignant phenotype of hypertrophic cardiomyopathy caused by Arg719Gln cardiac beta-myosin heavy-chain mutation in a Chinese family. *Clin. Chim. Acta* **310**, 131–139.
14. Mateu, E., Calafell, F., Lao, O., Bonne-Tamir, B., Kidd, J. R., Pakstis, A., et al. (2001) Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.* **68**, 103–117.
15. Shegogue, D., and Zheng, W. J. (2005) Applications note: capturing biological information with class-responsibility-collaboration cards. *Bioinformatics* **21**, 415–417.
16. MySQL Open Source DBMS. (2006) Url: http://mysql.org/.
17. JBoss Consortium. (2006) JBoss J2EE Web Server. Url: http://jboss.org/.
18. Apache Group. (2006) Apache HTTP Server 1.5. Url: http://apache.org/.
19. Apache Group. (2006) Apache Tomcat-Jakarta Java Web Server 5.0. Url: http://java.apache.org/.
20. Goodwill, J. (2001) *Developing Java Servlets 2nd ed.* Sams Publishing, Indianapolis IN, pp. 8–10.
21. Fung, D. C. -Y., Yu, B., Littlejohn, T., and Trent, R. J. (1999) An online locus-specific mutation database for familial hypertrophic cardiomyopathy. *Hum. Mutat.* **14**, 326–332.
22. Fung, D. C. -Y., Holland, E. A., Becker, T. M., Hayward, N. K., Bressac-de Paillerets, B., Melanoma Genetics Consortium, et al. (2004) *e*MelanoBase: an online locus-specific variant database for familial melanoma. *Hum. Mutat.* **21**, 2–7.
23. Stanke, F., Becker, T., Cuppens, H., Kumar, V., Cassiman, J. J., Jansen, S., et al. (2006) The TNFalpha receptor TNFRSF1A and genes encoding the amiloride-sensitive sodium channel ENaC as modulators in cystic fibrosis. *Hum. Genet.* **119**, 331–343.
24. EBI Mutation Database Recommendations. (2006) http://www.ebi.ac.uk/mutations/recommendations.
25. CIA's World Factbook may be found on http://www.cia.gov/cia/publications/factbook/index.html.

13

Protein Comparative Sequence Analysis and Computer Modeling

Brett D. Hambly, Cecily E. Oakley, and Piotr G. Fajer

Summary

A problem frequently encountered by the biological scientist is the identification of a previously unknown gene or protein sequence, where there are few or no clues as to the biochemical function, ligand specificity, gene regulation, protein–protein interactions, tissue specificity, cellular localization, developmental phase of activity, or biological role. Through the process of bioinformatics there are now many approaches for predicting answers to at least some of these questions, often then allowing the design of more insightful experiments to characterize more definitively the new protein.

Key Words: protein structure prediction; protein homology; BLAST; protein secondary structure; protein evolutionary relationships; protein function prediction.

Abbreviations: BLAST – basic local alignment search tool, E – expectation (value)

1. Introduction

Our starting point in protein characterization is the amino acid sequence, which has frequently been determined from a DNA sequence, but occasionally from direct protein sequencing. Importantly, sequence comparison is generally most effective when comparing the amino acid sequence, rather than the DNA sequence, since there is redundancy in the latter, i.e., for 20 amino acids and the stop signal there are 64 codons. Degeneracy in the DNA sequence comes from the codons' third bases and multiple codons coding for the same amino acid.

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

Comparison of protein sequence and/or structure is at present the most powerful approach for predicting the function of a protein. Interestingly, protein structure is frequently preserved more effectively than is sequence within protein families during evolution, although at present, sequence is generally more readily available than structure.

A plethora of online programs are available to assist in the analysis of protein sequences and structures. The two primary portals to these tools are either through the European-based Expasy site, or through the USA-sponsored NCBI Entrez site. However, many of the relevant programs that can be used at these sites are available in a range of pre-packaged programs, where multiple levels of analysis can be undertaken following a single data entry operation. Indeed the web-based services currently available are often ahead of the commercially produced individual computer packages. Many scientists are making their latest developments available to the wider scientific community prior

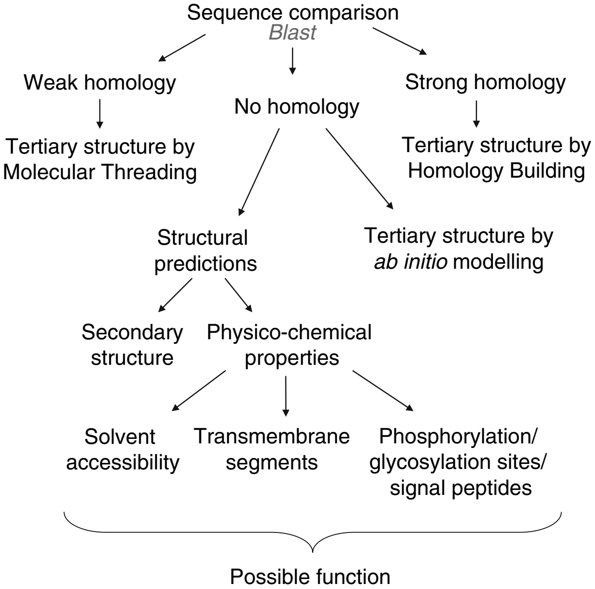

Sequence comparison
Blast
Weak homology
No homology
Strong homology
Tertiary structure by Molecular Threading
Tertiary structure by Homology Building
Structural predictions
Tertiary structure by ab initio modelling
Secondary structure
Physico-chemical properties
Solvent accessibility
Transmembrane segments
Phosphorylation/ glycosylation sites/ signal peptides
Possible function

Fig. 1. Summary of the elements that may be involved in the prediction of possible structural content and function of a protein product from its corresponding sequence. If strongly homologous proteins of known structure are readily found by sequence comparison, then tertiary structure, hence function, of the protein is likely to be readily predicted. Weaker or absent homology may require intermediate structural prediction, that may provide clues to the function of the protein.

to their commercialization, which usually targets the biotechnology and pharmaceutical industries.

Although many of the various analyses that can be undertaken are often able to be sourced in a pre-packaged form, allowing the work to be undertaken in a single operation, the individual elements of the analysis should still be considered **(Fig. 1)**. Broadly these operations can be divided into a number of stages:

1. Primary sequence similarity and alignment.
2. From sequence to protein secondary structure, protein topology and prediction of post-translation modifications.
3. Protein tertiary structure prediction.
4. Protein "functional" analysis.

2. Materials

2.1. Primary Sequence Similarity and Alignment

1. Input sequence is usually required in FASTA format. A sequence in FASTA format consists of an optional single-line description, followed by lines of sequence data. The first character of the description line is a greater-than (">") symbol in the first column. All lines should be shorter than 80 characters (ASCI file of single letter amino acids or nucleotides).
2. Access to BLAST online program:

   a. NCBI-BLAST website (www.ncbi.nlm.nih.gov/BLAST/)
   b. NCBI-BLAST tutorial (www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html)

2.2. From Sequence to Protein Secondary Structure, Protein Topology and Prediction of Post-Translation Modifications

1. The user can execute the required programs at a number of sites and web-accessible servers. Since the links and addresses of the hosting sites keep changing, the safest approach is to Google the names of the programs discussed below. There are also "super-sites" that provide services and links to most of the steps, e.g.,

   a. NCBI and Expasy servers (www.expasy.org),
   b. PredictProtein at Columbia University, (http://predictprotein.org/).
   c. Bioinfobank Institute in Poznan, Poland, (http://bioinfo.pl/meta/)
   d. BCM at Baylor College of Medicine (http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html)
   e. CMS Molecular Biology Resource at University of California San Diego (http://restools.sdsc.edu/)

2. These servers are continually updated and provide useful links to new programs and literature. Some of these servers will also act as Meta-servers and will submit the jobs to other relevant servers, translate the standard input into server-specific input and pass on the information from one stage to the next.
3. Structural domain databases can be found at
    a. ProDom (http://prodom.prabi.fr),
    b. SCOP (http://supfam.mrc-lmb.cam.ac.uk/superfamily/),
    c. Pfam (www.sanger.ac.uk/software/Pfam/)
4. Atomic level protein structures at PDB (www.rcsb.org).
5. Popular and reliable secondary structure prediction programs can be found at:
    a. PROF (http://predictprotein.org/)
    b. NNSSP (http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html)
    c. PREDATOR (http://bioweb.pasteur.fr/seqanal/interfaces/predator-simple.html)
6. A functional motifs database is found at PROSITE (www.expasy.org/prosite/).
7. Prediction of solvent accessibility, phosphorylation and glycosylation sites, signal peptides and transmembrane segments can be accomplished using links at the PredictProtein Meta-server.

2.3. Protein Tertiary Structure Prediction

Access to tertiary structure prediction programs are currently available at:

1. Modeller (http://salilab.org/modeller/)
2. 3D-JIGSAW (http://predicprotein.org/)
3. Swiss-model (http://predicprotein.org/)
4. PROCHECK (www. biochem.ucl.ac.uk/r̃oman/procheck/procheck.html)
5. 3DPSSM (www.sbg.bio.ic.ac.uk/∼3dpssm/index2.html)
6. Threader (http://bioinf.cs.ucl.ac.uk/threader/)
7. Rosetta (http://robetta.bakerlab.org/)

3. Methods

3.1. Primary Sequence Similarity and Alignment

Proteins with similar sequences are likely to have diverged from common ancestral genes, and are therefore likely to have similar structures and functions. While this is usually the case, there are exceptions, e.g., where structure may have been substantially preserved, but the sequences have diverged beyond recognition, or alternately, a sequence and/or structure may be substantially similar, but the function of the protein has evolved into something completely different.

To determine the degree of sequence homology, a database of known sequences is searched. The technology required to do this can reach a substantial level of complexity, but web-based programs can yield excellent results very rapidly. Sequence divergence results from single accumulated nitrogenous base changes, insertions or deletions. Generally, approximately 35% identity between sequences constitutes homology, but clear structural homology may exist with less than 5% sequence identity.

The basic local alignment search tool (BLAST) finds regions of local similarity between sequences and is the most popular bioinformatics program for this purpose. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches ***(1)***. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. The BLAST algorithm emphasizes speed over sensitivity. Speed is vital to making the algorithm practical for use on the huge genome databases that are available. BLAST is also often used as part of other algorithms that require approximate sequence matching.

1. Choose the program to use and the database to search: The program to be used will depend on the input sequence (nucleotide — blastn or protein — blastp) and the database will depend on the biological source of the sequence. Generally, the non-redundant (nr) database is the best starting point, since it covers almost all known sequences. Additional more specialized programs and databases are available, and are comprehensively listed and explained at the BLAST website. In particular, two programs may improve your detection of low homology, distantly related sequences: PSI-BLAST (Position Specific Iterative BLAST) detects weak homologs by building a profile from a multiple alignment of the highest scoring hits in an initial BLAST search, and PHI-BLAST (Pattern-Hit Initiated BLAST) combines matching of regular expressions with local alignments surrounding the match (*see* **Note 1** for further details of BLAST).
2. Input the data in FASTA format. Alternatively, a database accession number can be used if the sequence is already available in a database.
3. Set the program options or choose defaults. Some parameters worth considering are

   a. Generally gapped alignments are preferable since gaps frequently exist, even in relatively homologous sequences.
   b. If a previous search has yielded very large numbers of hits, then limiting the search to specific organisms may be helpful.
   c. The E (expectation) value measures the significance of particular hits. Values <0.1 generally represent significant hits, but values between 0.1 and 10, while not significant overall, may contain short segments of significant sequence homology, which may allow the tentative assignment of biochemical activities to the query sequence. The significance of any such regions must be

assessed on a case by case basis. Thus, an E value between 0.1 to 10 is usually appropriate, depending on the number of results generated.

d. Low-complexity regions, e.g., stretches of cysteine, hydrophobic regions in membrane proteins or coiled coils, tend to produce spurious, insignificant matches with sequences in the database which have the same kind of low-complexity regions, but are unrelated biologically. Checking this option will analyze the query sequence using the SEG program, and all amino acids in low-complexity regions will be replaced by X's which will appear in the alignment.

e. A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix," which assigns a score for aligning any possible pair of residues. In general, different substitution matrices are tailored to detecting similarities among sequences that are diverged by differing degrees. The BLOSUM matrix assigns a probability score for each position in an alignment that is based on the frequency with which that substitution is known to occur among consensus blocks within related proteins. The BLOSUM-62 matrix is among the best for detecting most weak protein similarities. For particularly long and weak alignments, the BLOSUM-45 matrix may be superior.

4. Set the output formatting options. These options may limit the number of hit sequences you obtain if set too low, but prevent excessive results if the search yields large numbers of homologues. The default options are a good start.
5. Perform the search. The results of the search are made available to you in several useful formats, either online or as an email.

3.2. From Sequence to Protein Secondary Structure, Protein Topology and Prediction of Post-Translation Modifications

Although evolutionary pressure to preserve protein function leads to a higher structural conservation than sequence conservation, sequence similarities can still be used to search for structural similarities, e.g., 25% residue identity for long sequences (>80 residues) will result in a similar structure (*Sander-Schneider relationship*) ***(2)***. To date, the structures of 36,000 proteins have been determined experimentally at atomic level resolution, mainly by either X-ray diffraction or NMR. The resulting library can be used to predict the structure of unknown proteins that share similar sequences. Since there are three basic secondary structures (α helices, β strands and loops of random coil) and a finite number of structural super families (approximately 1,500) our ability to make a structural prediction becomes a reasonable one.

Historically, the first generation of secondary structure predictions was based on single residue statistics, i.e., the propensity of a given amino acid to form an α helix, β strand or a loop ***(3)***. These methods resulted in an average accuracy of 55%. Second generation methods ***(4)*** introduced segment comparisons, i.e., the propensity of the central residue to form a particular secondary structure, taking into account the 8 adjacent residues, and an expanded database of protein structures, increasing the accuracy to about 60%. State-of-the-art, third-generation methods use evolutionary information contained in aligned multiple sequences of known structures ***(5)***. Currently these methods achieve an accuracy of 72%. If a given sequence shows a conformational preference in a number of structures, then it is likely to show a similar preference in the unknown sequence.

One of the most popular programs is PHD, and it is based on artificial neural networks ***(6)***. This method has dramatically improved the evaluation of β strands, which were a weak point with the previous approaches. Additionally, PHD evaluates the reliability of the prediction for any given residue, where the top third of "reliable" residues is predicted with 90% accuracy. Other programs of the third generation use a nearest-neighbor approach or long range interactions.

3.2.1. Overall Prediction Strategy

Irrespective of the algorithm used and differences in scoring functions, the basic strategy is to

1. Identify proteins with sequences that are similar, and align the unknown sequence with known sequences.
2. Predict the secondary structure.
3. Many physico-chemical and functional properties can be predicted at this stage:
   a. transmembrane fragments of the proteins,
   b. solvent accessibility of the residues,
   c. fortuitously in some cases, the function of the protein.

To achieve these predictions, open one of the Meta-servers, e.g., the PredictProtein portal, and then select the MetaPP option to allow choice of programs that are available.

3.2.2. Identify Proteins with Sequences that are Similar and Align the Unknown Sequence with Known Sequences.

The most popular database containing sequences is Swiss-Prot and TrEMBL. These sequence databases can be searched and aligned for homologous sequences using BLASTp or PSI-Blast, as described above. The PredictProtein site combines these databases into BIG, and uses the program MaxHOM to identify and align multiple sequences.

3.2.3. Predict Secondary Structure

Popular and reliable (>70%) third generation prediction programs are

a. PROF *(7)* (neural networks, improved PHD),
b. NNSSP (nearest neighbor approach) *(8)*, and
c. PREDATOR *(9)* (long range interactions, hydrogen bonds).

3.2.4. Physico-Chemical and Functional Characterization of Proteins:

A functional motifs database is found at PROSITE. Prediction of solvent accessibility, phosphorylation and glycosylation sites, signal peptides and transmembrane segments can be accomplished using links at the PredictProtein Meta-server.

3.3. Protein Tertiary Structure Prediction

Predicting the tertiary structure is far from a routine task. In a sense, we are trying to fold a polypeptide chain *in silico*, a very daunting challenge when one considers the number of possible conformations, and when this process occurs in cells it can often only be undertaken with the assistance of many proteins that help with folding. The approach to predicting tertiary structure is highly dependent on the extent to which homologous proteins of known structure can be identified.

3.3.1. Significant Sequence Identity—Homology Building

The process of tertiary structure prediction is simplified when the sequence identity is high. For example, if greater than 30% sequence identity exists then the structure of homologues differs in most cases by an average of less than 3 Å along the polypeptide chain. For these proteins, the new sequence is threaded into the polypeptide backbone of the known structure, the R-groups rotamers are chosen to minimize steric clashes and the overall energy. This process is called "homology" building. Popular homology programs are Modeller, 3D-JIGSAW and Swiss-model. Homology building is aided by the identification of structurally conserved regions in the family of homologous proteins, and multiple sequence alignment that identifies the gaps in the alignment to be filled by loops. Loops of 4 or 5 residues are modeled by either finding a template in the database or by energy minimization. Longer loops are more problematic. Often, the homology built models need to be energy optimized by Monte Carlo or molecular dynamics (simulated annealing) methods. The evaluation of the model is a must for any tertiary structure modeling attempts. The program PROCHECK evaluates dihedral angles of each residue and maps onto Ramachandran plots of commonly observed angles. For densely packed proteins, the accuracy is very high but one needs to be careful about the extended proteins where large rearrangements of the independent domains are difficult to predict.

3.3.2. Lower Sequence Identity

Lower sequence identity, e.g., 10–25%, calls for finding structurally conserved regions across proteins that are not homologous. There is a finite number of folds (600–1500) observed in proteins (the databases SCOP, CATH, and FFSP) and one can scan easily enough through all of them to fit a given sequence and its physical properties. For example, polar residues should match polar residues. If buried, the small side chains should substitute for other small side chains when they occur in the protein interior and so on. Threading programs, e.g. 3DPSSM or Threader, use a variety of scoring functions to evaluate whether or not the scanned sequence is likely to be occupying a given fold. The accuracy of the threading-based methods is around 3–10 Å.

3.3.3. No Sequence Identity—ab Initio Modeling

Ab initio modeling has to be used in cases when there are no recognizable sequence identities (<5%) or no recognizable folds in the databases. The evolutionary information is not detectable, and the structure prediction follows the physical folding pathway of a linear chain. Replicating the process *in silico* has many problems–a good analogy is that of finding an exit out of an inclined labyrinth. Following the path (energy) to guide the solution is only justified if the labyrinth is monotonously inclined, which is not the case in real life. Folding intermediates are not always of lower energy than their predecessors. If we cannot rely on the energy to guide each step we need the ability to sample all the possibilities, using both downwards and upwards steps, referred to as exhaustive sampling. Unfortunately, we do not have the means of assuring that sufficient sampling has taken place, nor can we be absolutely certain about the assumed energy functions. The problem with the latter is that even the smallest errors compound during folding. Nevertheless, a combination of folding with short segment patterns (ROSETTA, ***(10)***) or modeling of the folding on the lattice to constrain the number of possibilities shows considerable promise (*see* **Note 2**).

3.4. Protein Functional Analysis

Programs have been developed to determine the function of uncharacterized proteins directly from their sequence. An example of such a program is PROSITE, which utilizes a database of biologically significant sequences and sequence patterns associated with specific functional properties. From this, the program can rapidly identify to which known family of proteins (if any) the new sequence may belong. Frequently, the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment, but a section of the sequence may contain a particular cluster of residue types (a pattern or motif). These motifs arise because of particular requirements on the structure of specific region(s) of a protein which are important, e.g., for their binding properties or for their enzymatic activity. These requirements impose tight constraints on the evolution of those limited (in size) but important portion(s) of a protein sequence. Chapter 16 provides further discussion of how protein function is assessed.

4. Notes

1. BLAST is actually a family of programs (all included in the blastall executable) including

   Nucleotide-nucleotide BLAST (blastn): The input sequence is a DNA query, which returns the most similar DNA sequences from the DNA database.

   Protein-protein BLAST (blastp): The input sequence is a protein query, which returns the most similar protein sequences from the protein database.

   Position-Specific Iterative BLAST (PSI-BLAST): One of the more recent BLAST programs, this program is used for finding distant relatives of a protein. First, a list of all closely related proteins is created, which are then combined into a "profile" or "average" sequence. A query against the protein database is then run using this profile, and a larger group of proteins found. The process is then repeated through several iterations.

   Nucleotide 6-frame translation-protein (blastx): This compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

   Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx): It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.

   Protein-nucleotide 6-frame translation (tblastn): This program compares a protein query against the six-frame translations of a nucleotide sequence database.

   Megablast: For large numbers of queries *Megablast* speeds up the analysis of multiple sequences by concatenating many input sequences together to form a large sequence before searching the BLAST database, then post-analyzes the search results to glean individual alignments and statistical values.

2. The optimism surrounding tertiary structure prediction is to a large extent a consequence of the "blind" testing of the 3D prediction algorithms using structures (targets) whose conformation has been experimentally determined but not released to the scientific community. In a type of competition, every two years competing scientists submit their predictions of what these structures should look like, and an independent jury evaluates the agreement between the prediction and the actual structures (CASP competition (http://predictioncenter.org/casp7/)). The best predictions in the current competition, CASP7 (2006), are only a few Å root mean square deviation (rmsd) away from the actual structures. We should mention that one popular approach in this competition is to use automatic servers that will predict the secondary structure, determine the folds, align the sequences and model 3D structure by a variety of the above mentioned methods.

Acknowledgments

This work was supported by a grant from the American Heart Association.

References

1. 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.
2. Sander, C., and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68.
3. Chou, P. Y., and Fasman, G. D. (1974) Prediction of protein conformation. *Biochemistry* **13**, 222–245.
4. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.
5. Qian, N., and Sejnowski, T. J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884.
6. Rost, B., and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
7. Ouali, M., and King, R. D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.* **9**, 1162–1176.
8. Salamov, A. A., and Solovyev, V. V. (1997) Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **268**, 31–36.

9. Frishman, D., and Argos, P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng*. **9**, 133–142.
10. Misura, K. M., Chivian, D., Rohl, C. A., Kim, D. E., and Baker, D. (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl. Acad. Sci. U S A* **103**, 5361–5366.

14

Identification and Characterization of Microbial Proteins Using Peptide Mass Fingerprinting Strategies

Jonathan W. Arthur

Summary

Peptide mass fingerprinting is a simple, quick, cheap, and relatively effective method of identifying proteins from mass spectrometry data. Proteins extracted from the complex mixture comprising the proteome of a sample are individually digested with a proteolytic enzyme into a series of peptide fragments. The set of masses of these peptides, determined by mass spectrometry, form a peptide mass fingerprint of the protein. Comparison of this experimental fingerprint with the theoretical fingerprints of all known protein sequences for this organism, derived computationally from a protein sequence database, allows the identification of the particular protein. In this chapter, I discuss the technique including preparation for the peptide mass fingerprinting analysis, the appropriate selection of computational search parameters, and the analysis and interpretation of search results in the context of identifying proteins from microbial samples.

Key Words: peptide mass fingerprinting; proteomics; mass spectrometry; bioinformatics; data analysis.

Abbreviations: EST – expressed sequence tag; m/z – mass/charge; PMF – peptide mass fingerprinting

1. Introduction

Proteomics is rapidly emerging as an important discipline in life science research. The proteome is defined as the protein complement of the genome or, in other words, the entire set of proteins expressed by a genome. Proteomics,

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

it follows, is simply the study of proteomes. In the last few years, significant developments in proteomics technology have created the opportunity to rapidly and effectively separate, identify, and characterize hundreds or thousands of proteins expressed by a particular organism, in a particular tissue, at a particular time.

Proteomics is a broad discipline, incorporating a wide range of technologies and techniques in all aspects of the proteomic process. Despite this wide range of techniques, the method involving two-dimensional polyacrylamide gel electrophoresis (2D PAGE) coupled with matrix assisted laser desorption ionization mass spectrometry (MALDI-MS) and followed by peptide mass fingerprinting (PMF) for protein identification is still a very common technique. This is possibly due to this particular method of proteomic analysis involving relatively simple techniques while also being relatively inexpensive. Details of the 2D PAGE proteomic process can be found in many places including the seminal reference on proteomics edited by Wilkins et al. ***(1)***. Here, our focus is on PMF and how it can be used to identify and characterize proteins from microbial samples.

PMF was first described in 1993 by five separate research groups ***(2–6)***. As a result of the wet-laboratory part of the proteomic experiment prior to PMF, each spot on the gel, representing a unique protein derived from the original sample, has been digested with an enzyme such as trypsin to produce a series of peptides. The mass of each of these peptides has been measured on a mass spectrometer to produce a mass spectrum where each peak in the spectrum corresponds to the mass of one of the peptides derived from the protein being studied. This spectrum is thus a peptide mass fingerprint of the protein because the set of peptides, and thus spectral peaks, derived from each protein is expected to be unique.

The completion of various genome projects has resulted in knowledge of the complete genome sequence of various organisms and the identification of the entire set of open reading frames, and thus protein sequences, for particular organisms. In August 2006, there were 354 complete microbial genomes in the Entrez Genome resource provided by the National Centre for Biotechnology Information in the United States. Thus, it is possible to determine a theoretical peptide mass fingerprint for every protein potentially expressed by a particular organism. This is done by computationally searching the protein sequence for cleavage sites and cleaving the protein at this point to produce a peptide. The mass of the peptide is then calculated by adding the masses of the various residues, a N-terminal H atom, a C-terminal OH residue, and an additional

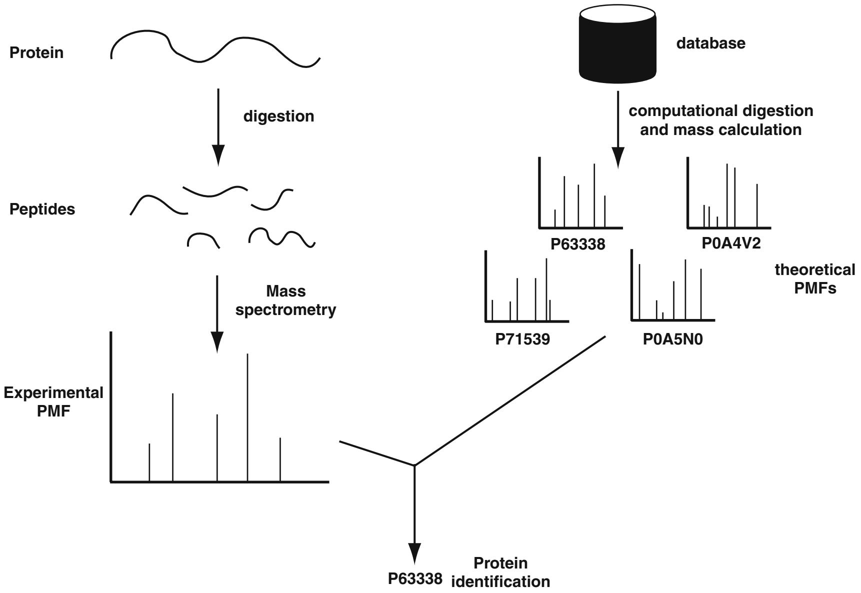


Fig. 1. Schematic drawing of the peptide mass fingerprinting process. The left hand side of the figure outlines the process for obtaining an experimental peptide mass fingerprint via enzymatic digestion of the protein and mass spectrometry. The right hand side of the figure outlines the process for obtaining a series of theoretical peptide mass fingerprints for each protein in the protein sequence database using computational methods. The protein is identified by comparing the experimental peptide mass fingerprint to each of the theoretical peptide mass fingerprints in order to find the best match.

H atom representing the charge attached to the peptide to match with the experimentally charged peptides.

Each of the theoretical peptide mass fingerprints is compared in turn to the experimental peptide mass fingerprint until a match is found. The match identifies the protein contained in this particular spot. The whole process can then be repeated for every spot on the gel. **Figure 1** shows an overview of the peptide mass fingerprinting process.

2. Materials

Before commencing a peptide mass fingerprinting analysis it is critical to prepare for the analysis by obtaining the requisite software, data, and associated information required for the analysis.

2.1. Choice of PMF Software Application

A range of software applications exist and are freely available for use over the internet in PMF. Most of the software applications are produced by companies who also sell versions of the software application, other related software packages, or services and support in the use of the software application. In many cases, the free, online version of the application is sufficient. It is usually only necessary to purchase the software if you have a specialized application, highly sensitive data you do not wish to transfer over the internet, or you wish to automate the process of identifying many proteins.

The most commonly used and freely available software applications for peptide mass fingerprinting are (1) Mascot, by MatrixScience ***(7)***, (2) Profound, by Genomic Solutions ***(8),*** (3) Phenyx, by GeneBio ***(9)***, and (4) Protein-Prospector, by University of California, San Francisco ***(10)***. The choice of PMF software is largely one of personal preference (*see* **Note 1**).

2.2. Species of Origin of the Sample

Identify, if possible, the species of origin of the sample. This is usually well known. PMF is not a particularly good technique for identifying the origin of an unknown sample. Also, clarify whether the sample is likely to be contaminated with proteins from another species. For instance, infected sputum may contain bacterial as well as human proteins.

2.3. Enzyme used to Digest the Proteins

Identify the enzyme used in the proteolytic cleavage of the proteins separated from the complex mixture of proteins comprising the proteome of the sample. In most cases this will be trypsin.

2.4. Artefactual Modifications

Identify all chemical treatments of the sample during the sample preparation process prior to the separation and array of proteins using 2D gel electrophoresis. Also identify whether the chemical treatments would have potentially modified certain amino acids in the proteins. For example, protein samples are usually reduced and alkylated to break disulphide bonds between cysteine residues in the protein and prevent these bonds from re-forming. This results in the addition of an alkyl chain to the cysteine residues, changing the mass of the cysteine residue, and thus the mass of the peptide containing this residue. This potential mass difference must be allowed for when identifying and characterizing the protein.

It is also necessary to identify the nature of the resultant modified residues. The two most common artefactual modifications are

1. The reduction and alkylation of cysteine residues with either acrylamide, resulting in cysteine residues with a propionamide modification, or iodoacetamide, resulting in cysteine modifications with a carbamidomethyl modification
2. Oxidation of methionine residues through exposure of the sample to air during the sample preparation process

2.5. Query Peak List

Finally, the most critical material for the analysis is the mass spectrum itself. Different mass spectrometers report the mass spectrum acquired from a gel spot in a variety of different data formats. In many cases, these formats are proprietary and often readable only by the software associated with the operation of the mass spectrometer. Even where the mass spectrum is produced in a human readable (often ASCII) data format (or can be converted to this format by the mass spectrometer software), the result is a pattern of changing signal intensity as a function of mass/charge (m/z) ratio. In contrast, most PMF software applications require a list of single m/z values for each peak in the spectrum, optionally with paired signal intensity information.

The process of converting a mass spectrum into a "processed spectrum", "stick spectrum", or "peak list" is called "peak picking" or "peak harvesting." Breen *et al.* ***(11,12)*** provide an example of one method of peak picking. Many mass spectrometers have associated software that will perform this operation.

The exact format of the data required will depend on the particular PMF software application. However, a single-column list of the m/z values for each peak in a regular text file allowing the list to be cut and pasted into the PMF software application will usually suffice. Some applications will accept comma or space separated data with two columns, the first containing the m/z value of each peak and the second containing the intensity of signal of the corresponding peak.

3. Methods

The methods described below outline the process of

1. undertaking a basic peptide mass fingerprinting search resulting in a list of potential identities for the protein ("hits"), and
2. interpreting the results of the search to identify the hit most likely to be the correct identity of this protein.

3.1. Identification of the Protein

This section describes the process of identifying a protein using PMF. In general, the process requires the user to input data in the form of a peak list derived from a mass spectrum produced from an enzymatic digest of a protein. This peak list will serve as input data to a PMF software application. A number of other user determined parameters are used to refine the search. The results of the analysis are then reported and interpreted by the user. Typically, a list of potential identifications ("hits") will be found and the user will interpret the information presented about each hit in order to determine which ones (if any) can confidently be chosen and thus determine the identity of the protein.

The method of entering data into the PMF software application, as well as the nature, extent, and use of the various user-defined parameters will vary between different applications. This section aims to be as specific as possible while maintaining a discussion that is generic enough to allow the steps to be applied to a variety of different PMF software applications.

3.1.1. Generic User Data

Some software applications require the user to submit a number of pieces of generic information such as your name, email address, a title for the search, a name for the sample being analyzed, etc. This information should be entered as required or desired.

3.1.2. Data Source

The data source or database indicates the repository of sequence information to be used during the PMF search. If a nucleotide database is chosen, the nucleotide sequences will be first translated to protein sequences. All the protein sequences in the database are then digested to provide theoretical PMFs to compare to the experimental data. The PMF software application will usually offer a selection of possible databases and the user chooses a single database from the list.

The selection of a database is largely a matter of personal preference. UniProt ***(13)*** is a good choice because it strikes a balance between completeness and quality (*see* **Note 2**). It provides a relatively non-redundant set of sequences about 10% of which form the SwissProt database and thus come with an extremely high quality manual annotation making characterization and biological interpretation of the role of the protein much easier.

3.1.3. Species and Taxonomy

Most PMF software applications allow the user to restrict the PMF search to a particular species or taxonomic grouping. Normally, the user will choose a species or taxonomic grouping from a list. To identify microbial proteins, choose the species of the organism you are working with. For example, if you are working with *Mycobacterium tuberculosis* then you would choose this option from the list of species. If your species is not specifically listed, choose the most relevant taxonomic grouping containing your species (*see* **Note 3**). Restricting the search to a particular species reduces the time taken to complete the search and eliminates false positive hits to proteins in other genomes through random matching of peptides.

3.1.4. Enzyme

It is necessary to indicate to the PMF software application the enzyme used to digest the protein in the wet-laboratory experiment. The software application will use cleavage rules corresponding to the chosen enzyme to determine the theoretical peptide masses. Trypsin is perhaps the most commonly used enzyme in preparing proteins for PMF.

3.1.5. Missed Cleavages

The user selects the number of missed cleavages the PMF software application should make allowance for. Missed cleavages result from incomplete digestion of the protein during the enzymatic digestion. This may occur because of inadequate time for digestion or amount of enzyme. More usually it represents cleavage points in regions of the protein poorly accessible by the enzyme. A value of zero or one is usually the most appropriate choice (*see* **Note 4**).

3.1.6. Modifications

There are two types of modifications to make allowance for: real modifications and artefactual modifications.

3.1.6.1. Artefactual

Artefactual modifications are post-translational modifications resulting from chemical treatment of the protein as part of the sample preparation process. These modifications must be allowed for in the PMF search.

Artefactual modifications are allowed for by selecting one or more modifications from the list provided in the software application. Some software applications provide two different ways of searching for modifications. The modifications can be applied to every potential site of modification (“fixed” modifications) or in a combinatorial fashion, checking for the possibility of zero, one, or more modifications on a peptide with multiple potential sites of modification (“variable” modifications). In this case, “fixed” modifications should be used in searching for the artefactual modifications (*see* **Note 5**).

Commonly, it is necessary to allow for the alkylation of cysteine residues and the oxidation of methionine residues.

3.1.6.2. Real

Real modifications are those post-translational modifications resulting naturally from the co- or post-translational processing of the protein in the organism from which it is derived. These include acetylation, deamidation, methylation, phosphorylation, and many others. These usually affect only a small number of peptides and potential modification sites in the protein sequence. As such, they are best searched for using the “variable” modifications option if available, preferably after the main search (*see* **Note 5**).

3.1.7. Mass and Isoelectric Point Filters

Some PMF software applications allow you to restrict the search to proteins in the database whose theoretical molecular weight and isoelectric point lie within a user-determined range. To do this, the user identifies the position of the protein spot on the gel and determines an experimental molecular weight and isoelectric point based on the location of the spot in the gel, often with reference to marker proteins or well-known proteins found elsewhere on the gel. An appropriate range containing this experimental location is then selected in the PMF software application. The experimental peak list is then compared only to the theoretical PMFs derived from proteins whose theoretical molecular weight and isoelectric point fall within this range. In general, an initial search should *not* be restricted by molecular weight or isoelectric point (*see* **Note 6**).

3.1.8. Error Tolerance

The error tolerance is the allowance made for experimental error in the measurement of the peptide masses on the mass spectrometer. It defines a range around the theoretical mass of a peptide where any experimental peak whose mass value falls within this range is determined to match the theoretical mass. The appropriate value of the error tolerance depends on the accuracy and precision of the mass spectrometer. A lower value reduces the number of false positive matches, so this value should be set as low as the quality of mass spectrum will allow. Typically, this should be less than 0.1 Da (*see* **Note 7**).

3.1.9. Minimum to Match

Some software applications allow you to specify a minimum number of matching peptides. This requires a potential protein hit to have at least this number of matching peptides before it will be considered as a potential hit. A typical value for an initial search is four, although many modern scoring systems make this parameter obsolete.

3.2. Interpreting the Result

Once the PMF search is completed, the PMF software application will generally return a set of results. These results usually need to be interpreted to identify the protein or proteins whose presence is indicated by the data in the mass spectrum. Different applications will present the results in different ways and make available different bioinformatic tools for visualizing and interpreting the results. In general, however, the results of a PMF search centre around a list of proteins (or "hits") that have one or more peptides with a mass matching one of the peaks in the experimental spectrum.

The following section describes a general method for interpreting the results of a PMF search.

3.2.1. Score

Most PMF search applications will calculate and assign a score to each hit. The score is usually the most reliable indicator for identifying the protein. Various different scoring mechanisms exist including MOWSE ***(3)***, probability based MOWSE ***(14)***, Bayesian posterior probability ***(15–17)***, and randomization distribution ***(18,19)***. In general, when using the score to identify a protein from a list of hits, a hit is selected as the most likely identification of the protein if the score is

1. Numerically high. Each different scoring mechanism will present a different range of scores. Thus the numerical value classed as "high" is determined from experience with a particular PMF software application. In general, it is not possible to compare raw scores between two different scoring mechanisms. For some scoring systems, it is not possible to compare raw scores between searches against two different databases and thus "high" must be interpreted according to what is high for other searches against the chosen database.
2. Noticeably higher than the score for other hits. Once again, "noticeably higher" depends on the range of scores generally given by the particular scoring algorithm
3. Statistically significant. Some scoring algorithms use statistical methods to define a score threshold separating statistically significant scores from those that are

insignificant. This makes it possible to separate those hits likely to be real hits from those resulting from random peptide matches.

3.2.2. Number of Matching Peptides

The number of matching peptides for a particular hit is also a useful guide for selecting a protein hit. In most cases, a high number of matching peptides corresponds to a real hit (*see* **Note 8**).

3.2.3. Coverage

Coverage is the percentage of residues in the protein hit amino acid sequence that was found in one of the matching peptides in the protein hit. For example, in the artificial case of a 10 amino acid protein with two tryptic peptides, one of length 6 amino acids and the other of length 4, one match to the longer peptide would result in 60% coverage.

The higher the coverage, the more confidence we can have in the hit being real. As a rule of thumb, a coverage greater than 1/n where n is the number of matching peptides provides strong evidence of a real hit. For example, with four matching peptides, coverage of 25% is a rough minimum required to accept a hit as real. As the number of matching peptides increases the estimated minimum coverage required decreases (*see* **Note 9**).

3.2.4. Intensity of Matching Peaks

In any mass spectrum there are usually high intensity peaks and low intensity peaks. A hit whose peptides match many of the high intensity peaks is stronger than one that does not. Matching high intensity peaks adds confidence to a protein identification. In contrast, a hit matching only low intensity peaks cannot be ruled out.

3.2.5. Gel Region Matching

The experimental molecular weight and isoelectric point for the protein can be determined from the position of the spot in the image of the gel. These values can be compared to the theoretical molecular weight and isoelectric point for the protein hit. A correlation between the two adds confidence to the identity. However, a mismatch does not necessarily detract from a particular identity. Post-translational modifications, truncations, and polymerizations can all change the experimental molecular weight and isoelectric point, leading to a mismatch with the theoretical values.

4. Notes

1. Most PMF software applications function in a relatively similar manner. The user enters a list of m/z values derived from the mass spectrum and selects a series of user-determined options. The most commonly used options are usually the same in all software applications. The user then examines the results, often with a set of small applications (tools) to aid in the interpretation of the result. Each software application does, however, have some unique features. These unique features may appeal to particular individuals or be useful in different situations, leading to a personal preference for a certain software application. It is, of course, possible to use more than one software application to undertake the peptide mass fingerprinting analysis. This is commonly practised in some laboratories as a way of bringing added confidence to the results.
2. For completeness, GenBank ***(20)***, or the regional equivalents, EMBL or DDBJ, may be the best option although redundancy in the database can make interpretation of the results more difficult. In some cases, the PMF software application will offer specialized databases for use in particular situations. For example, some PMF software applications allow you to search expressed sequence tag (EST) information (*see* **Chapter 17** for more discussion on ESTs). This can be useful if the species you are working with does not have many full gene or protein sequences in the database. A hit to an open reading frame in an EST sequence allows the user to use BLAST ***(21)*** or similar applications to identify a sequence homologue of the EST in another species, thus potentially identifying novel proteins during the PMF search (*see* **Chapter 13** for more discussion on BLAST).
3. There are two situations where you may want to search in a less specific taxonomic grouping. First, if you are working with tissue from another organism infected with bacteria, and it is important to identify both the bacterial proteins and the proteins from the other organism, you may wish to choose a taxonomic grouping broad enough to cover both organisms. For example, to identify proteins expressed in human sputum from individuals infected with *M. tuberculosis* you may wish to search against all species. Some PMF applications allow you to search two or more specific species enabling you to search *Homo sapiens* and *M. tuberculosis* simultaneously. You can also achieve the same results by conducting two separate searches. Second, if the species you are working with does not have many protein or nucleotide sequences in the database it may be necessary to search a broader taxonomic grouping containing one or more related species with completely sequenced genomes. For example, *Mycobacterium acapulcensis* has, at the time of writing, no sequences in UniProt. Thus, it will be impossible to identify proteins derived from this species using a search against UniProt and restricting the search to this species. However, by expanding the search to include all *Mycobacterium* species, it may be possible to identify homologues in the *M. tuberculosis* genome for *M. acapulcensis* proteins. Cross-species matching is of limited usefulness ***(22)*** because a single amino acid change (other than leucine to isoleucine or vice versa)

will change the mass of the peptide containing the residue and thus reduce the number of matching peptides. For this reason, the chosen taxonomic grouping should always be as specific as possible to the sample being analyzed.

4. In general, assuming good experimental technique, a value of zero should be used first. Once the identity of the protein is confirmed based on a search with zero missed cleavages, or at least there is good tentative evidence for a hit, the search can be repeated with one or more missed cleavages to see if the identification can be further confirmed or unmatched masses explained by the presence of missed cleaved peptides.
5. As artefactual modifications result from a chemical treatment they will usually affect most, if not all, potential modification sites in each peptide. By using "fixed" modifications, the time taken for the search is dramatically reduced because the software application only has to search for a mass matching the completely modified version of each peptide. The number of false positive matches is also dramatically reduced because the software application only considers the fully modified version of each peptide in the database and not all possible combinations. For example, the peptide MECAHCK would have an unmodified mass of 821.3103 Da. If the protein giving rise to this peptide was reduced and alkylated with acrylamide and exposed to oxidation, it could exist in up to six different forms (unmodified, methionine modified only, one cysteine modified only, two cysteines modified only, methionine and one cysteine modified only, methionine and two cysteines [i.e., fully] modified). Thus, searching with "variable" modifications greatly increases the chance of finding a matching peptide by chance leading to false positives. "Variable" modifications are most useful in secondary searches. Once the identity of the protein is confirmed based on a search with "fixed" modifications, or at least there is good tentative evidence for a hit, the search can be repeated with one or more "variable" modifications to see if the identification can be further confirmed or unmatched masses explained by the presence of modified peptides. Of course, this can also be done using alternative bioinformatics tools such as FindMod ***(23,24)***.
6. Restricting the search based on molecular weight and isoelectric point both improves the search speed (by reducing the number of theoretical peptide mass fingerprints needing to be checked) and reduces the number of false positives hits. On the other hand, it increases the possibility of a false negative. The form of the protein on the gel may have a distinctly different molecular weight or isoelectric point from the theoretical values calculated from the protein sequence in the database due to polymerization, truncation, splice variation, or post-translational modification. A restricted search is sometimes useful if you suspect you have a large number of false positives in the search results making it difficult to identify the protein. In this case, restricting the search may eliminate many of these false positives allowing a tentative true positive to be more easily detected.
7. Many software applications allow you to specify the error tolerance two ways. The first is a single value applied to the whole spectrum e.g. 0.1 Da. The second is

an error tolerance specified on a parts per million basis e.g. 100 ppm. The latter is more accurate as it allows for a larger error at larger mass values reflecting the way the mass spectrometer operates. A value of 100 ppm is equivalent to searching with an error tolerance of 0.1 Da in peptides with a mass of 1000 Da but an error tolerance of 0.05 Da in peptides with a mass of 500 Da.

8. The number of matching peptides needs to be interpreted with care. Large proteins generate a huge range of peptides of varying masses. As such, there is a much larger chance of finding matches to the experimental data by sheer chance in a large protein. For this reason, a large number of peptide matches in a large protein hit may not be indicative of a real hit.
9. In long proteins, it is also useful to examine the portions of the protein sequence covered by matching peptides. If the matching peptides are clustered at the start or the end of the sequence, it may indicate the protein whose identity is being determined has had a C- or N-terminal truncation leading to a modified version of the protein. Most PMF software applications have visualization tools for the coverage ("coverage maps") allowing easy visualization of any clustering of the peptide hits.

References

1. Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. (eds.) (1997) *Proteome Research: New Frontiers in Functional Genomics*. Springer-Verlag, Berlin.
2. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A* **90**, 5011–5015.
3. Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327–332.
4. Yates III, J. R., Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214**, 397–408.
5. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195**, 58–64.
6. Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338–345.
7. Mascot, http://www.matrixscience.com/ (2006) MatrixScience.
8. ProFound, http://65.219.84.5/service/prowl/profound.html (2006) Genomic Solutions.
9. Phenyx, http://www.phenyx-ms.com/ (2006) GeneBio.
10. ProteinProspector, http://prospector.ucsf.edu/ (2006) University of California San Francisco.

11. Breen, E. J., Hopwood, F. G., Williams, K. L., and Wilkins, M. R. (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* **21**, 2243–2251.
12. Breen, E. J., Holstein, W. L., Hopwood, F. G., Smith, P. E., Thomas, M. L., and Wilkins, M. R. (2003) Automated peak harvesting of MALDI-MS spectra for high throughput proteomics. *Spectroscopy* **17**, 579–596.
13. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159.
14. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.
15. Tang, C., Zhang, W., Fenyo, D., and Chait, B. (2000) Assessing the performance of different protein identification algorithms. *Proc. 48th ASMS Conf.*
16. Zhang, W., and Chait, B. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72**, 2482–2489.
17. Zhang, W., Tang, C., Fenyo, D., and Chait, B. (2000) A new method to evaluate the quality of database search results. *Proc. 48th ASMS Conf.*
18. Eriksson, J., Chait, B., and Fenyo, D. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* **72**, 999–1005.
19. Eriksson, J., and Fenyo, D. (2002) A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis. *Proteomics* **2**, 262–270.
20. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005) GenBank. *Nucleic Acids Res.* **33**, D34–D38.
21. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhao, Y., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
22. Wilkins, M. R., and Williams, K. L. (1997) Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J. Theor. Biol.* **186**, 7–15.
23. Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P. -A., et al. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289**, 645–657.
24. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005) Protein identification and analysis tools on the ExPASy Server, in *The Proteomics Protocols Handbook* (Walker, J. M. ed.), Humana Press, Totowa, NJ.

15

Statistical Analysis of Image Data Provided by Two-Dimensional Gel Electrophoresis for Discovery Proteomics

Ben Crossett, Alistair V. G. Edwards, Melanie Y. White, and Stuart J. Cordwell

Summary

Standardized methods for the solubilization of proteins prior to proteomics analyses incorporating two-dimensional gel electrophoresis (2-DE) are essential for providing reproducible data that can be subjected to rigorous statistical interrogation for comparative studies investigating disease-genesis. In this chapter, we discuss the imaging and image analysis of proteins separated by 2-DE, in the context of determining protein abundance alterations related to a change in biochemical or biophysical conditions. We then describe the principles behind 2-DE gel statistical analysis, including subtraction of background noise, spot detection, gel matching, spot quantitation for data comparison, and statistical requirements to create meaningful gel data sets. We also emphasize the need to develop reproducible and robust protocols for protein sample preparation and 2-DE itself.

Key Words: Image analysis, two-dimensional gel electrophoresis, data analysis, proteomics.

Abbreviation: 2-DE – two-dimensional gel electrophoresis

1. Introduction

Two-dimensional gel electrophoresis (2-DE) remains a technology of choice in the proteomics era ***(1)***, despite having recently celebrated its 30th birthday ***(2)***. The method is based on the separation of complex, or pre-fractionated, mixtures

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

of proteins using isoelectric focusing in the first dimension and sodium-dodecyl sulfate - polyacrylamide gel electrophoresis (SDS-PAGE) in the second. The resulting gels contain a profile of proteins represented as individual, or multiple, "spots" on the stained image. 2-DE remains a powerful technology since it provides a visual tool for monitoring global changes in protein expression or abundance. Many hundreds to thousands of individual proteins can be separated simultaneously on the two-dimensional gel matrix. Even more importantly, 2-DE provides a means of discerning post-translational modifications, that result in the alteration of the x,y-co-ordinates of a protein-of-interest. Alternative technologies, including multi-dimensional liquid chromatography coupled to tandem mass spectrometry of complex peptide mixtures ***(3)***, even with the advent of highly sensitive mass tags for comparative quantitation ***(4,5)***, are still unable to discern readily subtle changes in protein modification without the use of specialized affinity approaches, while protein cleavage remains only detectable using 2-DE ***(6)***.

The proteomics approach is generally used to determine how cells respond to a change in their environment, both surroundings (chemical and nutrient) and genetic (gene knock-out or over-expression). The use of 2-DE to understand these changes relies on two major factors to provide confidence in the acquired data ***(7)***: (1) biological replicates and (2) gel replicates. Typically, a large number of 2-DE gels are required to satisfy statistical criteria for protein "spot" changes between data sets. These overcome the issue of gel-gel variations (both a result of protein sample preparation discrepancies and of the 2-DE gel-running process) and limit non-specific biological differences caused by variation between individuals (particularly where animal tissue samples are used) ***(8–10)***. Furthermore, the use of large gel data sets allows rigorous statistical analysis of spot changes to be undertaken. Standard deviation on normalized spot densities should be considered prior to calculation of the *n*-fold change between gels produced from the separation of proteins expressed under varied conditions.

The question remains as to what 2-DE gel data actually represent? Many published manuscripts discuss the changes seen in 2-DE gel comparisons using terms such as "expression" and / or "up or down-regulation." Such terminology is not strictly correct. Visible e.g. silver stains or Coomassie blue, and fluorescent dyes actually measure protein "abundance" at a given point in time. Abundance is a function not only of the expression of a gene/protein at the transcriptional/translational level, but is also dependent on the protein life-span, or half-life. Therefore, a protein that is the product of a gene that is being actively transcribed (and therefore is well-*expressed*), but has a very short half-life, would not be an abundant, heavily stained protein on 2-DE gels. Expression of proteins is best measured using radiolabeling of actively translated protein. The use of radiolabels and fluorescent dyes in a combined approach to overcome this problem has been undertaken by the group of Michael Hecker and colleagues in Greifswald, Germany. This group has termed the phrase "dual-channel color imaging" of 2-DE gels. Two gels are analyzed; the first is generated by pulse radiolabeling using ^{35}S methionine (expression measurement), while the second gel is stained with a visible or fluorescent dye (abundance measurement). The spots on each gel are then given a color specific to the method of spot visualization, and the images are overlaid to determine which proteins are highly expressed under a given condition versus those that are abundant. The gel imaging software that is used to perform the analysis of these gel sets has been commercialized by DeCodon ***(11)***.

Since large amounts of data are acquired over multiple 2-DE gel replicates, data analysis by computer software is inevitable. Visual inspection of gels, however, remains an important and often under-rated first method for several reasons—first, such inspection allows the user to make a judgment on the quality of the gel images and whether any meaningful data can be achieved by using them; second, the eye remains a powerful imaging tool capable of detecting both very obvious and more subtle changes in protein abundance. Large replicate gel sets make visual inspection for the production of statistical analysis impossible, however, and many groups have developed computer algorithms to assist in 2-DE gel analysis ***(12–14)***. These algorithms are now almost all commercially available – e.g. PD-Quest, Progenesis, z3, DeCodon and Melanie, among others (*see* **Note 1**).

This chapter describes how 2-DE gel data sets are analyzed to achieve statistically meaningful data. This incorporates some discussion of how to choose a data set as well as some opinions on how to utilize the data. We discuss the use of image analysis software to compare 2-DE gel sets. Although the principles of image analysis are practically identical between software packages, we illustrate our discussions with basic methods compared using PD-Quest software.

2. Materials

Many commercial software packages are currently available. An excellent summary of the types of image supported, relative costs of purchase and compatible platforms is given by Raman and colleagues ***(12)*** (*see* **Note 2**). Choice of software package must be made in consideration of the number of gels that are expected to be run over the lifetime of the product, the relative cost of the software and the customer support available. Our view is that laboratories interested in image analysis packages should generate a high quality series of gels (3 control and 3 test) and use these to trial several programs prior to purchase.

3. Methods

3.1. Gel Quality

The ability to acquire meaningful data from comparisons using 2-DE relies on the use of high quality gel data to begin with. The expression "rubbish in, rubbish out" holds true for 2-DE image analysis (*see* **Note 3**). Undertaking a laborious computer-intensive analysis on poor quality data will only result in false positives (allocation of protein spots as being significantly altered following a change in environment that does not actually reflect a real change, including artifacts of the gel running process).

3.1.1. How Many Gels?

The experiment needs to be sufficiently powered such that the number of biological replicates is statistically viable. More samples are needed for studies involving individual-to-individual variation, e.g., animal tissue studies, than are needed for batch cultured cells, e.g., cell culture or microbiological species. To some extent, the number of biological replicates will dictate the number of gel replicates. For each protein sample preparation, it is preferable to run at least 3 gels, to account for gel-to-gel irreproducibility. A generally acceptable "rule-of-thumb" is that the more high-quality replicate 2-DE gels that are run (*see* **Note 4**), the more accurate and robust the resulting statistical data (*see* **Notes 3–5**).

3.2. Manual Analysis of 2-DE Gel Images

Visual inspection of 2-DE gel images from "control" and "test" groups is important to detect gross differences in spot patterns (**Fig. 1**) associated with the test conditions under study (*see* **Note 6**). This step serves two real purposes; first, to determine whether each individual gel is of sufficient quality to merit inclusion in the image analysis gel set; and second, to determine whether there are obvious differences in the gel patterns, and therefore that the "test" conditions have influenced gene / protein expression or protein abundance. A fair rule of thumb should be that if no very obvious differences can be detected by eye then the "test" may not be sufficient to induce changes, or that 2-DE technology is of insufficient sub-cellular specificity and/or sensitivity to

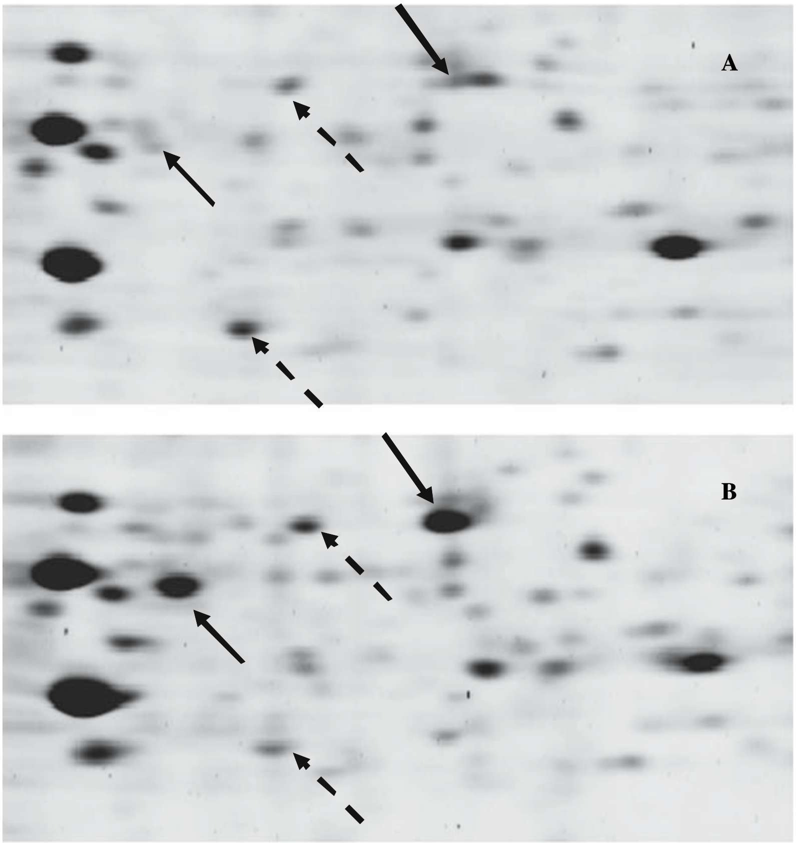


Fig. 1. Visual comparison of 2-DE gels **(A)** and **(B)**. High-quality, reproducible gels allow easy detection of protein spot appearances or disappearances (arrows), as well as even subtle changes in protein abundance (broken arrows).

detect meaningful changes in lower abundance proteins outside the scope of gel analysis.

3.3. Spot Detection

Image analysis of 2-DE gels for comparative purposes relies on the detection of spots *(16)*. This generally means determining the outlines of the separated spots such that spot densities (based on staining intensities) can be calculated. This task is not as immediately obvious as it appears. The first consideration is to remove any artifacts of the staining process. These appear as "speckles"

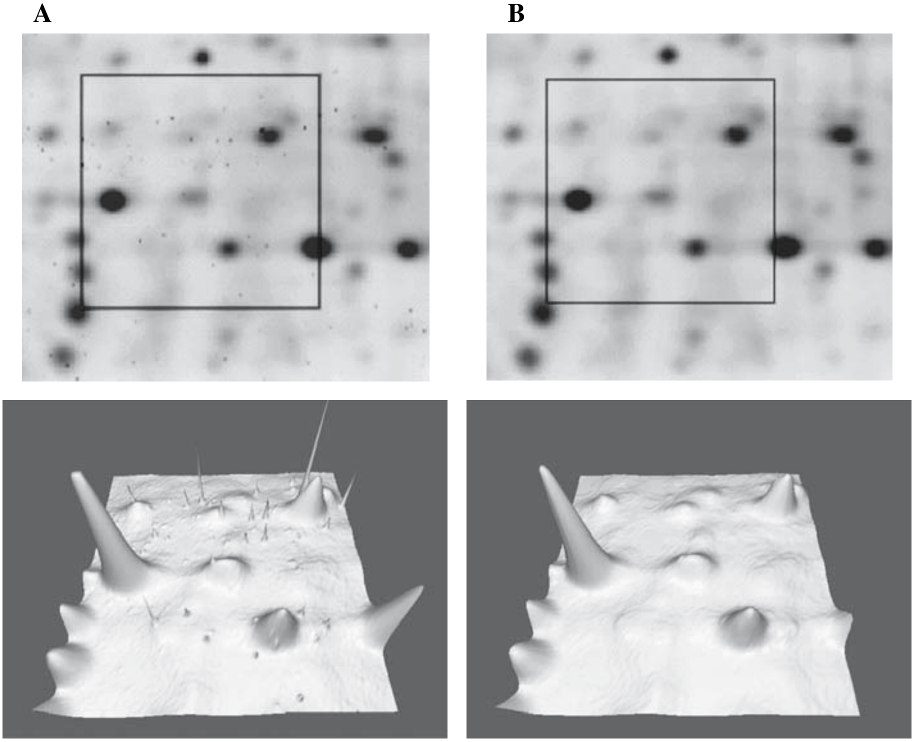


Fig. 2. Removal of artifacts from 2-DE gel images. Panel **(A)** shows 2-DE gel with “speckles” resulting from fluorescent staining. Panel **(B)** shows the same 2-DE gel image after filtering. Lower panels show 3-D Viewer representation, before and after filtration.

on the gel image and can be removed by setting the minimum spot diameter to a level greater than the diameter of the speckles (**Fig. 2**). Many packages include a “3-D Viewer” (**Fig. 2**), and these allow for the ready detection of spot artifacts. Other considerations include cropping the gel image so that the edges of the gels are removed and other artifacts, such as the sample application point, are removed. The second aspect to spot detection is far more complex – detecting those that do not separate into discrete spots. Proteins that separate into heavily staining “streaks” or lightly staining smears (especially where this is reproducible rather than a result of the 2-DE process itself) remain difficult to detect correctly ***(17)***.

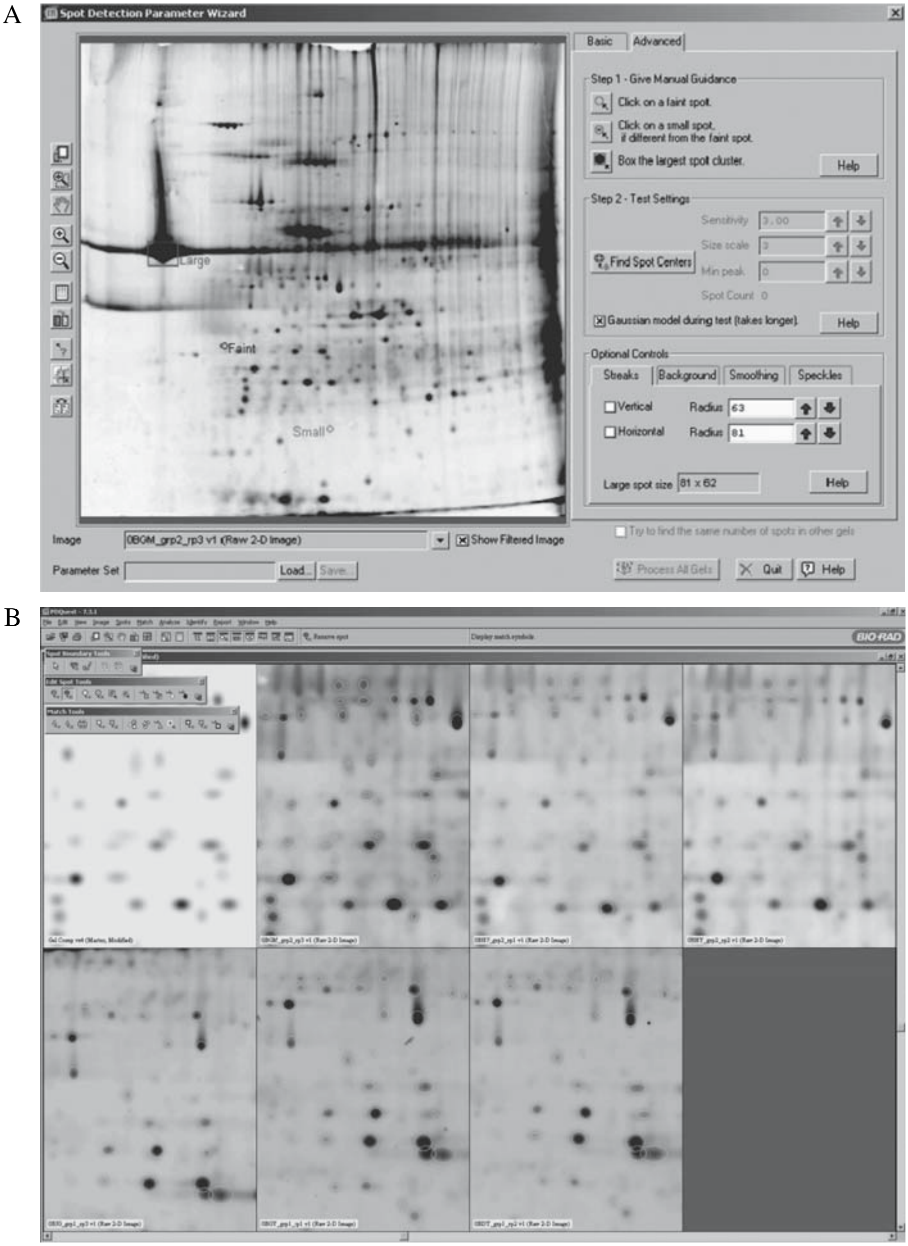
The process of detecting spots is generally automated, but requires significant user input to register the spots correctly. Initially, the user selects lightly stained, small spots and heavily stained, large spots to provide the limits of detection (**Fig. 3a**). The software then takes these criteria and attempts to detect the remaining spots on the gel (**Fig. 3b**). The user can then add additional spots to update the criteria. It is likely that several spots will need to be manually defined.

For large 2-DE gel data sets, some programs use "master" and "slave" gels. These are composites of all the individual gels within each set. To create the composite gel requires a decision on the reproducibility quality of the individual gels. That is, how many times should a spot appear on individual gels to be counted on the "master" gel. The more stringent this value, the higher the gel quality, but at the expense of the number of spots that can be compared between gel sets.

1. Gels should be scanned in high resolution and formatted as .tiff files (.gsc files for PD-Quest where using Quantity One software and a Bio-Rad imager such as a Molecular Imager Fx).
2. Crop the gel by removing the area above the highest mass marker, and the area below the lowest mass marker, as well as the areas on the left and right hand sides of the gel, including the molecular mass marker lane. Use the "Advanced crop" tool to ensure all cropped gels are the same.
3. Filter the image using the "Filter Wizard" to remove gel stain artifacts (check "salt" and "pepper" and set to 3 × 3. If there are still speckles on the resulting filtered image, re-check the setting to 5 × 5) (*see* **Note 7**).
4. Detect spots using the "Spot Detection Wizard." Use the normal settings as a default and then select the following: "Gaussian" distribution, remove vertical and horizontal streaks. After the first round of detection the values can be manually adjusted to optimize spot detection. Once optimized, the settings can be used for a series of gels, if they are similar enough, or re-optimized for each individual gel.

3.4. Gel–Gel Matching

Once the spots are detected, the next and final task is to "match" the corresponding spots between the gels, so that comparison of the data can be performed. The gel matching process can also be a laborious task, since gels become warped (expansion and contraction caused by the various stains and washes, and the length of time spent in each solution) and therefore the "master" and "slave" images cannot simply be overlapped (**Fig. 4**) *(18)*. In the matching process, the user may stipulate some landmarks on each gel and then allow the software to match the corresponding spots in that particular region (*see* **Note 8**). The more landmarks that are added, the better the spot matching, as



Basic | Advanced

Step 1 - Give Manual Guidance
Click on a faint spot.
Click on a small spot, if different from the faint spot.
Box the largest spot cluster.
Help

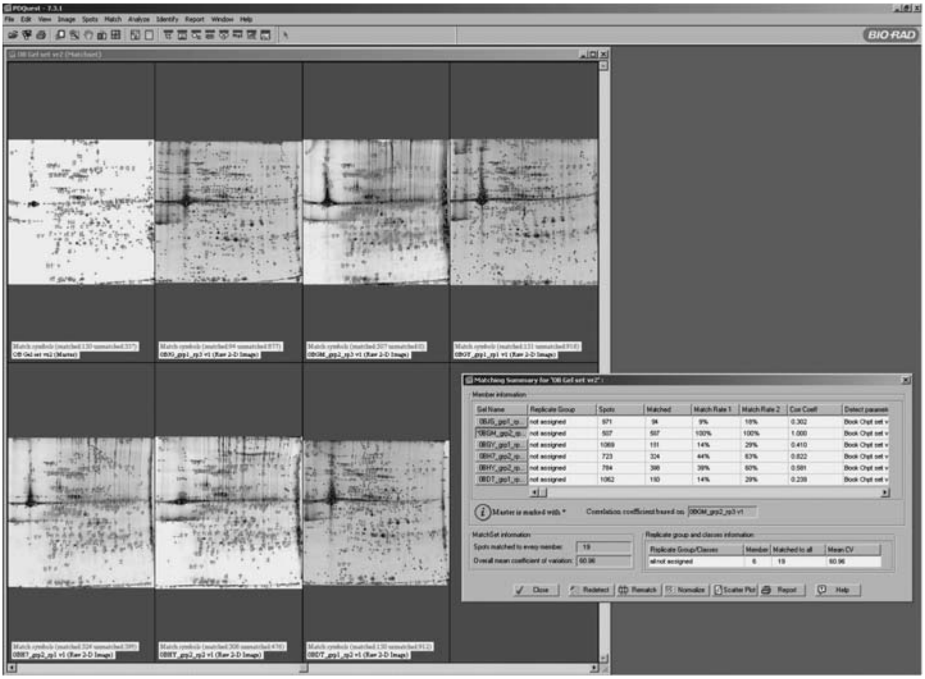
Step 2 - Test Settings
Find Spot Centers
Sensitivity 3.00
Size scale 3
Min peak 0
Spot Count 0
☒ Gaussian model during test (takes longer).
Help

Optional Controls
Streaks | Background | Smoothing | Speckles
☐ Vertical Radius 63
☐ Horizontal Radius 81
Large spot size 81 x 62
Help

☐ Try to find the same number of spots in other gels
Process All Gels | Quit | Help

Image 0BGM_grp2_rp3 v1 (Raw 2-D Image) ☒ Show Filtered Image
Parameter Set Load... Save...

Fig. 3. Spot detection on 2-DE gels. **A:** User selected criteria for spot definition in PD-Quest. **B:** Spots detected based on those criteria (white circles).


PDQuest - 7.3.1
File Edit View Image Spots Match Analyze Identify Report Window Help
BIO-RAD
Matching Summary for '08 Gel set ve2'
Close
Redetect
Rematch
Normalize
Scatter Plot
Report
Help

Member information

Gel Name	Replicate Group	Spots	Matched	Match Rate 1	Match Rate 2	Cor Coeff	Detect paramet
0BJG_grp1_rp...	not assigned	971	94	9%	18%	0.302	Book Chpt set v
*0BGM_grp2_rp...	not assigned	507	507	100%	100%	1.000	Book Chpt set v
0BGY_grp1_rp...	not assigned	1069	151	14%	29%	0.410	Book Chpt set v
0BH7_grp2_rp...	not assigned	723	324	44%	63%	0.822	Book Chpt set v
0BHY_grp2_rp...	not assigned	784	306	39%	60%	0.581	Book Chpt set v
0BDT_grp1_rp...	not assigned	1062	150	14%	29%	0.239	Book Chpt set v

Master is marked with * Correlation coefficient based on: 0BGM_grp2_rp3 v1

MatchSet information

Spots matched to every member: 19

Overall mean coefficient of variation: 60.96

Replicate group and classes information

Replicate Group/Classes	Member	Matched to all	Mean CV
all not assigned	6	19	60.96

Fig. 4. Spot matching between gels from two different data sets (control and test). **Upper panel** shows 4 gels (master on the left and three individual gels) comprising the control set; **lower panel** of 3 gels comprises the test set. Grey shows spots that are matched; black shows spots that have not yet been matched. This process, even with high-quality 2-DE gels can be time-intensive.

much of the warping may be regional, particularly at the extremes of the gel where expansion and contraction may be most noticeable (*see* **Note 9**).

5. Gel-gel comparisons are performed in “Matched sets.” These can either be created during the batch processing of gels or afterwards by clicking on Match > New Match Set.
6. To assess the quality of the matching it is possible to switch between various overlays which each display different useful information. Place the mouse over the gel of interest and press the following keys:

   a. Spot crosshairs—F5
   b. Show ellipses—AltF5
   c. Vector offset—shift F7
   d. Matched (green letters) and unmatched spots—F8

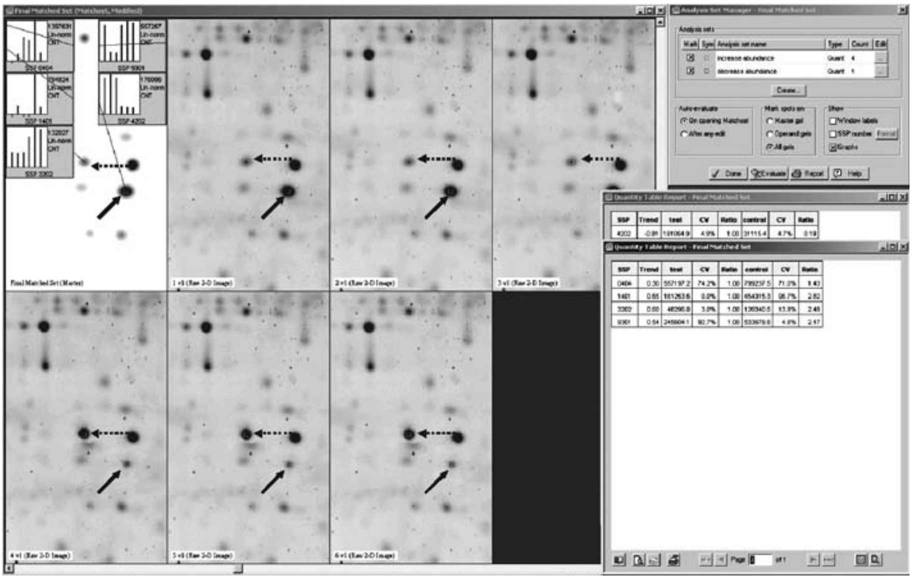
e. Only unmatched spots—Shift + F8
f. Remove all—Esc.

It can also be useful to look at other parameters such as the “Image Stack Tool” (which allows the user to flick between multiple gel images) or the “Scatter Plot Tool”, both of which are on the “Analyze” menu.

3.5. Spot Quantitation and Comparison

Spots that are least conserved between gels (compared via spot density alone) tend to be those at the extremes of abundance—the very large, dense proteins spots and the very low level, faint spots. This is because these spots are most sensitive to the irreproducibilities associated with 2-DE gels. Large spots tend to focus poorly and may have both vertical and horizontal streaks associated with them, furthermore, the image analysis programs will often detect them as multiple smaller spots, rather than as a single, large spot. When very abundant, intensely staining spots are measured they may “saturate” the density reading, again providing a lack of reproducibility. Very faint spots may appear to be altered under test conditions, simply because even minor differences in the concentration of protein added to the first dimension gel will lead to sizable quantitative differences. Therefore, nearly all programs use some type of “normalization”—the use of protein spots that do not change as internal controls to calibrate all the remaining data. The choice of spots that normalize is itself problematic as it must be certain that the spot actually does not change under the test conditions, even subtly, and thus bias the remaining data. Amongst a large gel set, however, the spots of “least change” are generally employed.

7. For comparison of groups of gels, gels need to be assigned to replicate groups by using the “Create Replicate Groups” tool available through the Analyze menu. All the individual gels within each group should undergo spot detection prior to inclusion in the group.
8. When all the gels have been assigned to a replicate group, the “Group Consensus” tool (from within the “Analyze” menu) can be used to check that as many of the spots have been matched as possible (*see* **Note 10**). This tool can determine whether there is consensus within a group, or between control and test groups, thus allowing for quantitation and comparison.
9. Once as many spots have been matched as possible, the “Analysis Set Manager” (accessed via the “Analyze” menu) to highlight the similarities and differences. Using this tool it is possible to build up a set of quantitative and qualitative comparisons, which can then be grouped further using Boolean analysis to add or subtract members of each group from each other.



Analysis Set Manager - Final Matched Set

Analysis sets

Mark	Sym	Analysis set name	Type	Count	Edit
☒	□	increase abundance	Quant	4	...
☒	□	decrease abundance	Quant	1	...

Create...

Auto-evaluate: ● On opening Matchset ○ After any edit

Mark spots on: ○ Master gel ○ Operand gels ● All gels

Show: □ Window labels □ SSP number ☒ Graphs

Done | Evaluate | Report | Help

Quantity Table Report - Final Matched Set

SSP	Trend	test	CV	Ratio	control	CV	Ratio
4202	-0.81	101064.9	4.9%	1.00	31115.4	4.7%	0.19

Quantity Table Report - Final Matched Set

SSP	Trend	test	CV	Ratio	control	CV	Ratio
0404	0.30	557197.2	74.2%	1.00	799237.5	71.0%	1.43
1401	0.65	161263.6	0.0%	1.00	654315.3	96.7%	2.82
3302	0.60	48296.8	3.0%	1.00	136940.5	13.9%	2.48
9301	0.54	245604.1	92.7%	1.00	533679.6	4.9%	2.17

Fig. 5. Comparison of spot quantities between 2-DE gel sets. Spots elevated in abundance are shown by broken arrows, those decreased in abundance by full arrows. Spot comparisons on each of six individual gels are shown on master gel **(top left)**. Statistical analyses are shown in the table on the right side of the image; SSP, spot number; Ratio, up- or down- *n*-fold change.

3.6. Statistical Analysis

2-DE gel data must be statistically analyzed (**Fig. 5**). Each individual spot change should be represented by three statistics: (1) mean spot density of each individual spot averaged over the number of gels analyzed; (2) standard error (standard error of the mean or standard deviation) of those mean spot densities (*see* **Note 11**); (3) Students' t-test to determine the statistical significance of the change; and (4) the *n*-fold change – a + or – change determined by dividing the mean spot density for a spot of interest from the "test" conditions divided by the mean spot density for the same spot from the "control" conditions (*see* **Notes 6 and 12**).

4. Notes

1. Several authors have attempted to compare software packages ***(12,14)***. Most conclude that some features are better in one program versus another, but that no single program is consistently better in every aspect than all the others. Therefore,

choice of software will be based on price, technical support, and user preferences. It should be noted that at least one study has suggested that variance in the analysis of 2-DE gel sets may be due to subtleties in the image analysis algorithm employed ***(15)***.

2. There are many commercially available image analysis options. Most vendors provide 21–30-day trials of "limited" or "fully functional" versions of their software, often available through the vendor website.
3. Many 2-DE gel users have difficulty in defining "good" from "bad" gels. This tends to be a highly subjective process. However, proceeding to image analysis on poor quality gels will only result in poor quality assumptions about protein differences between control and test conditions. There are some rules about defining gel quality: (1.) Discard gels where the majority of spots appear as streaks (either horizontal or vertical)—streaks provide very poor quality data for image analysis. (2.) Discard gels where the spots are hazy or poorly focused. (3.) Discard gels where the distribution of spots is not even in the SDS-PAGE dimension.
4. Under no circumstances should data from a single gel set ever be considered for publication. Many journals, including *Proteomics*, now set minimum standards for the publication of 2-DE gel data which should be considered ***(19)***.
5. Gels of "lower" quality, but not discarded gels, need much greater statistical analysis; that is, an increase in both biological and gel replicates to provide a meaningful data set. We have found that for bacterial samples (where gel quality is generally high), replicate biological preparations run on triplicate gels (6 *high-quality* gels per control or test group) is generally sufficient ***(20)***. However, for complex mammalian tissue samples ***(21)***, where gel quality is often poorer we have used up to 18 biological replicates and triplicate gels (54 *good-quality* gels). This also reflects the need to account statistically for individual variation between human or animal tissues. Finally, the use of more biological and gel replicates allows the user to work with lower statistically significant *n*-fold changes (for example, ±1.5-fold, rather than 2.0-fold).
6. Spot differences between control and test groups measured by 2-DE gels are generally defined by their *n*-fold change. This literally means the fold difference between the spot volume/density/pixel value (often in ppm) averaged across all the gels of the control group and all the gels of the test group. The generally accepted, minimum fold difference value is ±2.0-fold, which means an increase or decrease of 100% of the control spot quantity. This is clearly biased towards lower abundance proteins, which have a lower spot density and hence any change may be viewed as significant. Therefore, we often suggest that in any test condition, at least some major spots should be altered to consider the test biologically relevant. Increased numbers of biological and gel replicates may allow this value to go as low as ±1.5-fold change (50% increase or decrease in the mean spot density in test compared to control gels). Lower *n*-fold difference values may be based on gel-gel variance, particularly where few replicates have been performed.

7. Use of the “Filter Wizard” is optional, but should be used where gels have been stained with fluorescent dyes, such as Sypro Ruby *(22)*, which leave small “speckles” on the surface of the gel. The manufacturers of Deep Purple™ Stain *(23)*, suggest that no speckles remain on the surface of gels stained with this dye.
8. Some algorithms require the user to determine spot landmarks between “reference” and “test” gels. These “landmarks” define a spot viewed in the reference gel as the same x,y-coordinates as a spot in the test gel. The region of the two gel sets surrounding that landmark will then be warped to fit the best image match. More landmarks that are chosen between the two gels provide better accuracy in the spot matching process, particularly where the gels themselves are warped due to osmotic effects, or idiosyncrasies of the gel pouring and IEF gel positioning.
9. The use of very strict settings in the gel matching process require that a spot is in exactly the same location relative to its land-marked neighbors and will provide less matches, whereas less exact criteria provide more tolerance with the problem that false positive matches may occur.
10. No two gels sets will ever have 100% of spot matches correct. There is always a trade-off between a large number of correct matches and false positives. Generally, the higher the number of correct matches, the higher the number of false positives. Only manual data interpretation will overcome these issues. In real terms, most researchers will identify the “most” significant matches by visual inspection of the gels, and use the image analysis to determine statistical data on those spot differences.
11. Standard error of the mean for spot densities should be within the range of ±10–20%. Wider variances than this suggest irreproducibility of the 2-DE gel data.
12. It is wise to employ a spot volume cut-off when considering the results from spot matching and quantitation. Otherwise, a bias toward very low abundance spots being detected as statistically significantly altered between control and test conditions may occur.

References

1. Görg, A., Weiss, W., and Dunn, M. J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4**, 3665–3685.
2. O’Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
3. Washburn, M. P., Wolters, D., and Yates III, J. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.
4. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999.

5. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386.
6. White, M. Y., Cordwell, S. J., McCarron, H. C. K., Tchen, A. S., Hambly, B. D., and Jeremy, R. W. (2003) Modifications of myosin-regulatory light chain correlate with function of stunned myocardium. *J. Mol. Cell. Cardiol.* 35, 833–840.
7. Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M. C. M., Görg, A., Hecker, M., et al. (2006) Guidelines for the next 10 years of proteomics. *Proteomics* **6**, 4–8.
8. Choe, L. H., Aggarwal, K., Franck, Z., and Lee, K. H. (2005) A comparison of the consistency of proteome quantitation using two-dimensional electrophoresis and shotgun isobaric tagging in *Escherichia coli* cells. *Electrophoresis* **26**, 2437–2449.
9. Challapalli, K. K., Zabel, C., Schuchhardt, J., Kaindl, A. M., Klose, J., and Herzel, H. (2004) High reproducibility of large-gel two-dimensional electrophoresis. *Electrophoresis* **25**, 3040–3047.
10. Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T., and VanBogelen, R. A. (2003) Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **3**, 1912–1919.
11. Bernhardt, J., Buttner, K., Scharf, C., and Hecker, M. (1999) *Electrophoresis* **20**, 2225–2240.
12. Raman, B., Cheung, A., and Marten, M. R. (2002) Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. *Electrophoresis* **23**, 2194–2202.
13. Marengo, E., Robotti, E., Antonucci, F., Cecconi, D., Campostrini, N., and Righetti, P.G. (2005) Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems. *Proteomics* **5**, 654–666.
14. Rosengren, A. T., Salmi, J. M., Aittokallio, T., Westerholm, J., Lahesmaa, R., Nyman, T. A, et al. (2003) Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels. *Proteomics* **3**, 1936–1946.
15. Wheelock, A. M., and Buckpitt, A. R. (2005) Software-induced variance in two-dimensional gel electrophoresis image analysis. *Electrophoresis* **26**, 4508–4520.
16. Rogers, M., Graham, J., and Tonge, R. P. (2003) Using statistical image models for objective evaluation of spot detection in two-dimensional gels. *Proteomics* **3**, 879–886.
17. Rogers, M., Graham, J., and Tonge, R. P. (2003) Statistical models of shape for the analysis of protein spots in two-dimensional gel images. *Proteomics* **3**, 887–896.
18. Voss, T., and Haberl, P. (2000) Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: consequences for comprehensive data analysis. *Electrophoresis* **21**, 3345–3350.

19. http://www3.interscience.wiley.com/homepages/76510741/2120_instruc.pdf.
20. Cordwell, S. J., Larsen, M. R., Cole, R. T., and Walsh, B. J. (2002) Comparative proteomics of *Staphylococcus aureus* and the response of methicillin-resistant and methicillin-sensitive strains to Triton X-100. *Microbiology* **148**, 2765–2781.
21. White, M. Y., Cordwell, S. J., McCarron, H. C. K., Prasan, A. M., Craft, G., Hambly, B. D., et al. (2005) Proteomics of ischemia/reperfusion injury in rabbit myocardium reveals alterations to proteins of essential functional systems. *Proteomics* **5**, 1395–1410.
22. Lopez, M. F., Berggren, K., Chernokalskaya, E., Lazarev, A., Robinson, M., and Patton, W. F. (2000) A comparison of silver stain and SYPRO Ruby Protein Gel Stain with respect to protein detection in two-dimensional gels and identification by peptide mass profiling. *Electrophoresis* **21**, 3673–3683.
23. Mackintosh, J. A., Choi, H. Y., Bae, S. H., Veal, D. A., Bell, P. J., Ferrari, B. C., et al. (2003) A fluorescent natural product for ultra sensitive detection of proteins in one-dimensional and two-dimensional gel electrophoresis. *Proteomics* **3**, 2273–2288.

16

Online Resources for the Molecular Contextualization of Disease

Chi N. I. Pang and Marc R. Wilkins

Summary

Searching online resources can provide medical researchers with an efficient means of gathering existing knowledge on the molecular causes of disease. The researcher may choose to explore the following areas, e.g., genetic mutations associated with the disease, function and cellular sub-localization of the associated protein(s) and their protein interaction partners. Using a small case study, examining the disease retinoblastoma, this chapter guides the reader through the relevant information contained within relevant databases. It is shown that the integration of online biological knowledge with genomic and proteomic experimental data provides insights into the understanding of diseases in their molecular context.

Key Words: protein–protein interactions, disease, gene ontology, subcellular localization.

Abbreviations: GO – gene ontology; HPRD – Human Protein Reference Database; PTM – post-translational modification; RB1 – retinoblastoma

1. Introduction

Functional genomics and proteomics approaches are becoming increasingly widely used for the investigation, and understanding of disease. Microarray or proteomic techniques are frequently used for the comparison of gene or protein expression between diseased and control tissues, resulting in the identification of genes or proteins that are aberrantly expressed. At this point, investigators typically ask a series of fundamental questions. These include

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

1. Is the function of the protein known?
2. Is the protein known to be associated with disease?
3. Are mutations already described for this gene or protein?
4. Is the subcellular location and/or tissue distribution of the protein known? With increasing studies of protein-protein interactions and their association with disease, the researcher may also wish to ask:
5. Does the protein interact with any others?

This chapter will introduce a number of online resources which the medical researcher can use to contextualize their results from microarray and proteomics experiments. For the sake of clarity, we will write this chapter as a small case study, examining the *RB1* gene and protein, associated with the disease retinoblastoma. Whilst this protein was not discovered using microarray or proteomics experiments, instead having been discovered through classical genetics and molecular biology ***(1)***, we note that the application of proteomics/microarray experiments have shown the RB1 protein to be under-expressed in association with the disease ***(2)***.

2. Methods

There are two starting points that are useful for the contextualization of a gene or protein of interest. These are either with the name of the gene or protein, or with the name of the disease itself. Here we assume that the starting point will be with a gene or protein of interest. Each section is organized under a subheading, being a question or questions which researchers may choose to ask.

2.1. Is the Function of the RB1 Protein Known? Is the Protein Associated with Disease? What Mutations are Known?

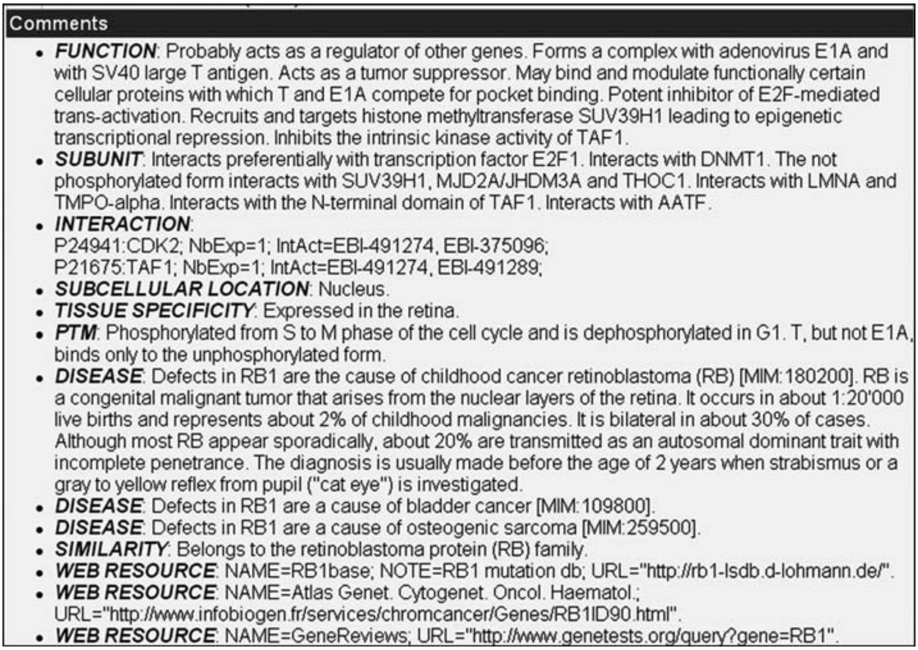
2.1.1. Searching the Swiss-Prot Database

Swiss-Prot is a curated protein sequence database. It provides useful annotation, such as the protein's function, cellular sub-localization, post-translational modifications and amino acid sequence variants. The entry for each protein provides hypertext links to a comprehensive diversity of useful biological databases ***(3,4)***.

The Expasy website (http://ca.expasy.org/) allows the user to search for a protein by entering the protein or gene name. For the retinoblastoma-associated protein, the user would need to type "retinoblastoma-associated" into the search field on the top of the Expasy home page. It then provides the user a list of proteins to select from. Selecting the appropriate protein name from the results list leads the user to the full Swiss-Prot entry for the retinoblastoma-associated protein, accession number P06400, entry name RB_HUMAN.

There are several parts of the database that are of interest to the reader, notably the comments field and the feature table. The comments field (**Fig. 1**) provides the reader with information on the protein's function, subunit information if part of a complex, protein-protein interactions, cellular sub-localization, tissue specificity, post-translational modification (PTM) and any diseases associated with the protein. The feature table (**Fig. 2**) provides information on amino acid variants of the protein. The "VARIANTS" field gives the position of the amino acid variant in a protein, what amino acid it is mutated to, and the associated disease. The feature table also contains information on the location and type of PTM in the "MOD_RES" field. Information on the enzyme that catalyses the addition of a PTM may also be found here. Note, however that fatty acid modifications are not recorded as "MOD_RES" but "LIPID," glycosylation is recorded as "CARBOHYD," molecular cross-linking as "CROSSLNK" and disulfide bonds as "DISULFID."

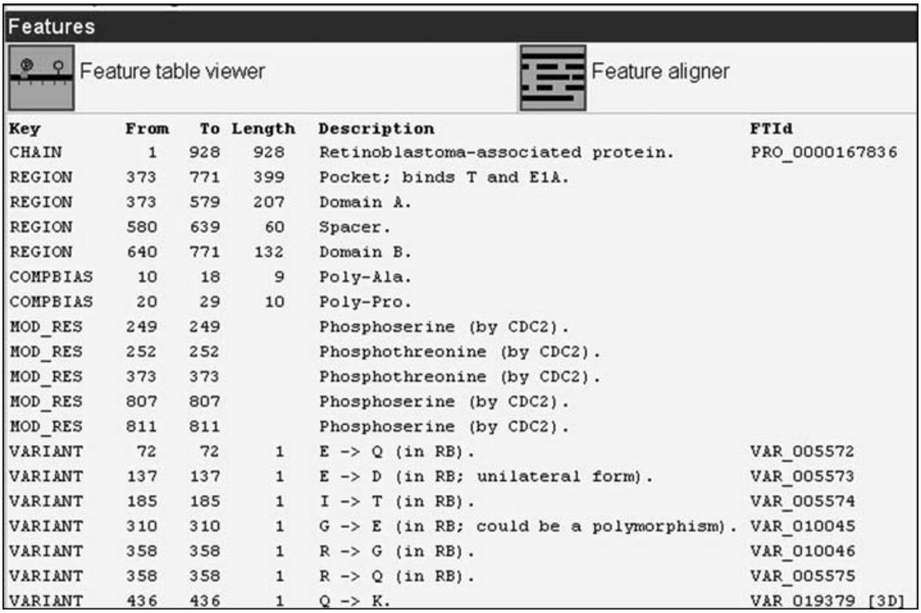
There are four classes of PTM reliability in Swiss-Prot. The first and most reliable class concerns modifications that have strong experimental



Comments

- ***FUNCTION***: Probably acts as a regulator of other genes. Forms a complex with adenovirus E1A and with SV40 large T antigen. Acts as a tumor suppressor. May bind and modulate functionally certain cellular proteins with which T and E1A compete for pocket binding. Potent inhibitor of E2F-mediated trans-activation. Recruits and targets histone methyltransferase SUV39H1 leading to epigenetic transcriptional repression. Inhibits the intrinsic kinase activity of TAF1.
- ***SUBUNIT***: Interacts preferentially with transcription factor E2F1. Interacts with DNMT1. The not phosphorylated form interacts with SUV39H1, MJD2A/JHDM3A and THOC1. Interacts with LMNA and TMPO-alpha. Interacts with the N-terminal domain of TAF1. Interacts with AATF.
- ***INTERACTION***:
P24941:CDK2; NbExp=1; IntAct=EBI-491274, EBI-375096;
P21675:TAF1; NbExp=1; IntAct=EBI-491274, EBI-491289;
- ***SUBCELLULAR LOCATION***: Nucleus.
- ***TISSUE SPECIFICITY***: Expressed in the retina.
- ***PTM***: Phosphorylated from S to M phase of the cell cycle and is dephosphorylated in G1. T, but not E1A, binds only to the unphosphorylated form.
- ***DISEASE***: Defects in RB1 are the cause of childhood cancer retinoblastoma (RB) [MIM:180200]. RB is a congenital malignant tumor that arises from the nuclear layers of the retina. It occurs in about 1:20'000 live births and represents about 2% of childhood malignancies. It is bilateral in about 30% of cases. Although most RB appear sporadically, about 20% are transmitted as an autosomal dominant trait with incomplete penetrance. The diagnosis is usually made before the age of 2 years when strabismus or a gray to yellow reflex from pupil ("cat eye") is investigated.
- ***DISEASE***: Defects in RB1 are a cause of bladder cancer [MIM:109800].
- ***DISEASE***: Defects in RB1 are a cause of osteogenic sarcoma [MIM:259500].
- ***SIMILARITY***: Belongs to the retinoblastoma protein (RB) family.
- ***WEB RESOURCE***: NAME=RB1base; NOTE=RB1 mutation db; URL="http://rb1-lsdb.d-lohmann.de/".
- ***WEB RESOURCE***: NAME=Atlas Genet. Cytogenet. Oncol. Haematol.; URL="http://www.infobiogen.fr/services/chromcancer/Genes/RB1ID90.html".
- ***WEB RESOURCE***: NAME=GeneReviews; URL="http://www.genetests.org/query?gene=RB1".

Fig. 1. Comments field for the retinoblastoma-associated protein from the Swiss-Prot database. The accession number for the RB1 protein is P06400.



Key	From	To	Length	Description	FTId
CHAIN	1	928	928	Retinoblastoma-associated protein.	PRO_0000167836
REGION	373	771	399	Pocket; binds T and E1A.	
REGION	373	579	207	Domain A.	
REGION	580	639	60	Spacer.	
REGION	640	771	132	Domain B.	
COMPBIAS	10	18	9	Poly-Ala.	
COMPBIAS	20	29	10	Poly-Pro.	
MOD_RES	249	249		Phosphoserine (by CDC2).	
MOD_RES	252	252		Phosphothreonine (by CDC2).	
MOD_RES	373	373		Phosphothreonine (by CDC2).	
MOD_RES	807	807		Phosphoserine (by CDC2).	
MOD_RES	811	811		Phosphoserine (by CDC2).	
VARIANT	72	72	1	E -> Q (in RB).	VAR_005572
VARIANT	137	137	1	E -> D (in RB; unilateral form).	VAR_005573
VARIANT	185	185	1	I -> T (in RB).	VAR_005574
VARIANT	310	310	1	G -> E (in RB; could be a polymorphism).	VAR_010045
VARIANT	358	358	1	R -> G (in RB).	VAR_010046
VARIANT	358	358	1	R -> Q (in RB).	VAR_005575
VARIANT	436	436	1	Q -> K.	VAR_019379 [3D]

Fig. 2. Excerpt of the feature table for the retinoblastoma-associated from the Swiss-Prot database. The accession number for the RB1 protein is P06400.

evidence. A further three classes concern PTMs which have not been experimentally verified. The most reliable of these are modifications inferred by taxonomic similarity, and are labeled “by similarity.” PTM information labeled as “probable” have some experimental evidence and should be found in the native protein. Modifications that have been predicted only by protein sequence analysis tools are denoted as “potential.”

In the case of RB1, the protein is annotated in Swiss-Prot as a tumor suppressor, which interacts preferentially with transcription factor E2F1. It is also known to interact with CDK2 and TAF1 proteins. The protein is localized in the nucleus, and may be phosphorylated in five different amino acid positions. The RB1 protein is associated with diseases such as the childhood cancer retinoblastoma, bladder cancer and osteogenic cancer. RB1 is also involved in pinealoma; this is described in the OMIM database discussed later.

2.1.2. Searching the Online Mendelian Inheritance in Man Database

The Online Mendelian Inheritance in Man (OMIM) is a database of genetic disorders. It provides links to literature in the PubMed literature database, to gene sequences in the NCBI database and other data sources. It is a resource used primarily by physicians and medical practitioners concerned with genetic disorders and genetics researchers ***(5)***.

The OMIM website allows the user to search for a disease by typing multiple keywords in the search box at top of the home page (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). In the case of RB1, the keyword "retinoblastoma" is sufficient. The identification number for the retinoblastoma database entry in OMIM is 180200. There are a number of sections of each OMIM entry that are of particular interest. They include the brief description of the disease, its gene map locus, molecular genetics, pathogenesis, gene function and allelic variants.

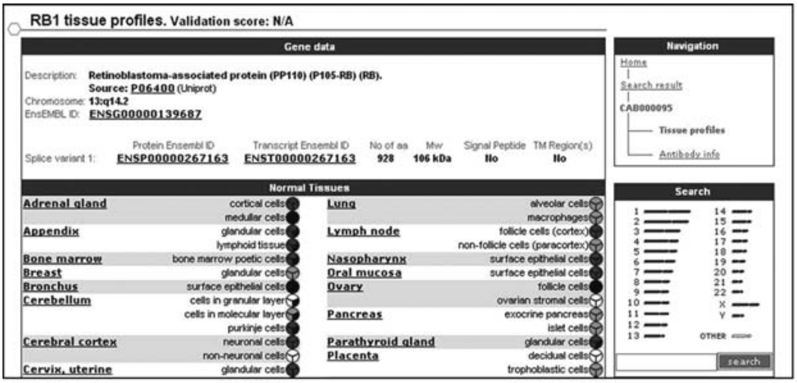
For RB1, it is recorded in the gene function section that RB1 is modulated by phosphorylation and dephosphorylation during different stages of the cell cycle. The RB1 protein is unphosphorylated in the G0/G1 phases, but it is mostly phosphorylated during S and G2 phases (***6–8***). The allelic variant entries include information on how the nucleic acid mutations affect the gene product.

2.2. Where is the Protein Found in the Body and Inside the Cell?

2.2.1. Searching the Human Protein Atlas

The human protein atlas (www.proteinatlas.org) is an ongoing project that aims to map the localization and expression of every human protein in all tissues of the body. The protein atlas database includes a large number of images of normal tissues and a variety of disease states (*see* **Note 1**). These are taken from immunohistochemically stained tissue sections, generated by use of specific antibodies generated against different antigens in the body. A brown-black color in an image highlights the location where the antibody has bound to its corresponding antigen. A blue stain is used for visualizing microscopic features in the same tissue samples. It may stain both cellular and extracellular materials. Each tissue type is represented by three images from unique patients. For most cancer tissue types, duplicate samples from 12 or more patients have been recorded, although some only have duplicate samples from 4 patients. All images have been analyzed by specialized image analysis software and validated by expert histopathologists. The reliability of each image is recorded in the database (*see* **Note 2**) ***(9)***.

To search for the retinoblastoma-associated protein in the human protein atlas, "RB1" or "retinoblastoma" is typed into the search field on the atlas home page (www.proteinatlas.org). In the search results page, click on the link under the "Antibody ID" column corresponding to the *RB1* gene name. This brings up an overview of RB1 localization throughout the body (**Fig. 3**). The overview



RB1 tissue profiles. Validation score: N/A

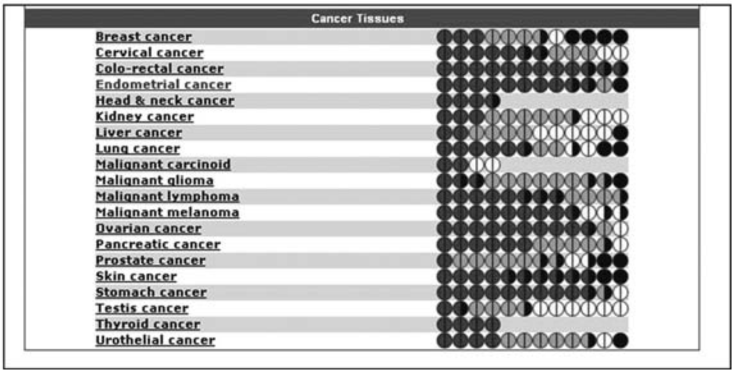
Gene data

Description: Retinoblastoma-associated protein (PP110) (P105-RB) (RB).
Source: P06400 (Uniprot)
Chromosome: 13:q14.2
EnsEMBL ID: ENSG00000139687

	Protein Ensembl ID	Transcript Ensembl ID	No of aa	Mw	Signal Peptide	TM Region(s)
Splice variant 1:	ENSP00000267163	ENST00000267163	928	106 kDa	No	No

Normal Tissues

Adrenal gland	cortical cells	Lung	alveolar cells
	medullar cells		macrophages
Appendix	glandular cells	Lymph node	follicle cells (cortex)
	lymphoid tissue		non-follicle cells (paracortex)
Bone marrow	bone marrow poetic cells	Nasopharynx	surface epithelial cells
Breast	glandular cells	Oral mucosa	surface epithelial cells
Bronchus	surface epithelial cells	Ovary	follicle cells
Cerebellum	cells in granular layer		ovarian stromal cells
	cells in molecular layer	Pancreas	exocrine pancreas
	purkinje cells		islet cells
Cerebral cortex	neuronal cells	Parathyroid gland	glandular cells
	non-neuronal cells	Placenta	decidual cells
Cervix, uterine	glandular cells		trophoblastic cells

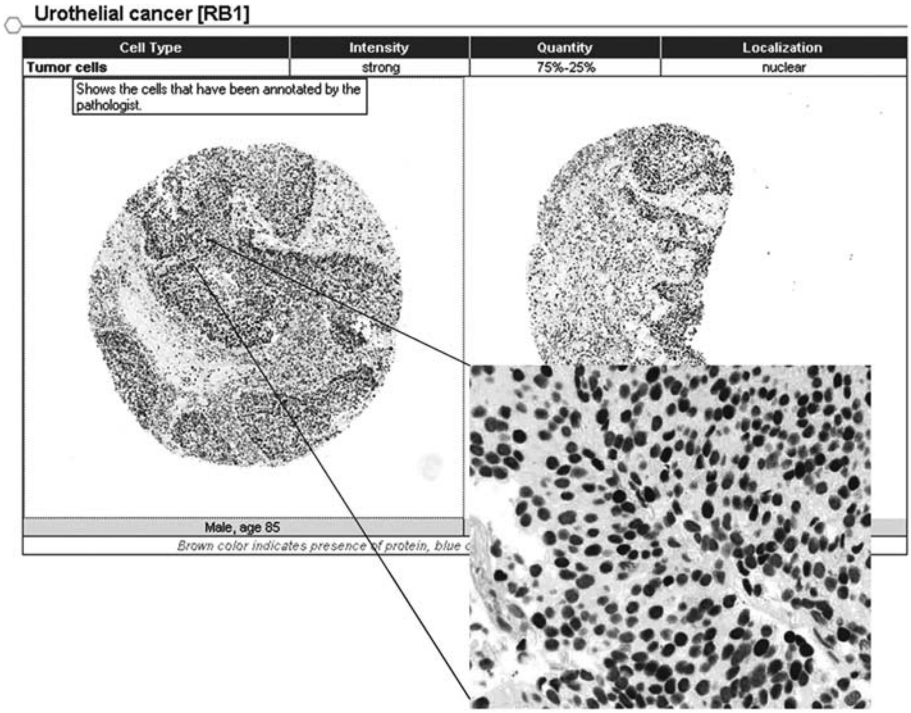


Cancer Tissues

Breast cancer
Cervical cancer
Colo-rectal cancer
Endometrial cancer
Head & neck cancer
Kidney cancer
Liver cancer
Lung cancer
Malignant carcinoid
Malignant glioma
Malignant lymphoma
Malignant melanoma
Ovarian cancer
Pancreatic cancer
Prostate cancer
Skin cancer
Stomach cancer
Testis cancer
Thyroid cancer
Urothelial cancer

Fig. 3. Tissue distribution of the RB1 protein from the human protein atlas. The top panel shows where the protein is found in normal tissues, and the bottom panel shows where the tissue is found in cancer tissues. The page can be accessed at www.proteinatlas.org/tissue_profile.php?antibody_id=95

includes an "annotation summary", describing the cellular sub-localization information for different tissues. For RB1, it showed strong and distinct staining in the nucleus of all tissue types, whether they were normal or malignant tissues. Fibroblasts, inflammatory cells and neuronal cells were also stained in the nuclei. In the protein atlas, users have to select images for tissue types most relevant to the disease under investigation. For example (**Fig. 4**), mutation in the retinoblastoma protein would cause urothelial carcinoma, commonly called bladder cancer. Normal bladder tissue and urothelial cancer tissue would be appropriate to view for this investigation.



Urothelial cancer [RB1]

Cell Type	Intensity	Quantity	Localization
Tumor cells	strong	75%-25%	nuclear

Shows the cells that have been annotated by the pathologist.

Male, age 85

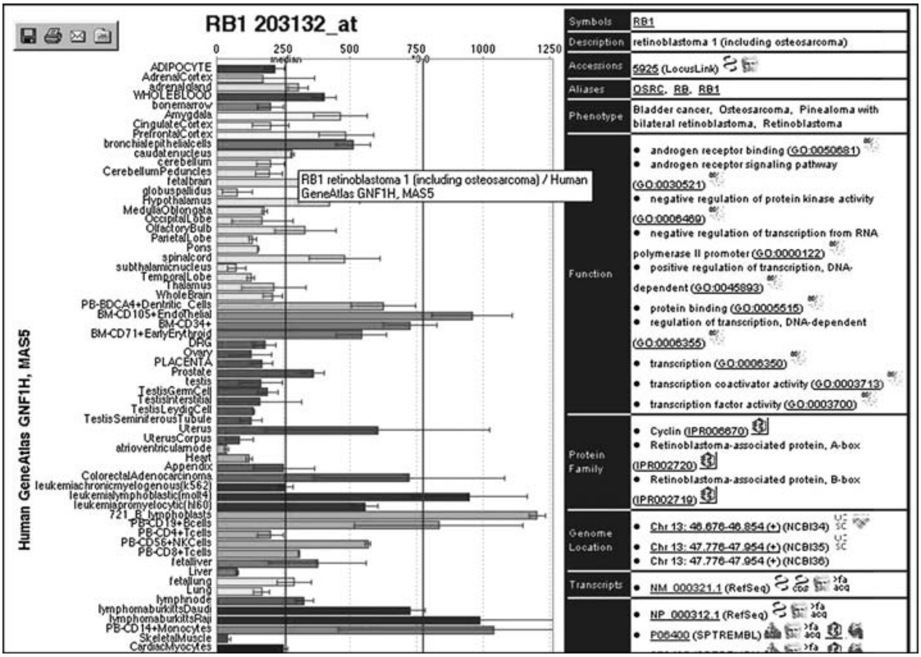
Brown color indicates presence of protein, blue c

Fig. 4. Histological view of RB1 protein expression in uorthelial cancer. The expanded view can be accessed by clicking on the lower resolution image. This page can be accessed at www.hpr.se/cancer_unit.php?antibody_id=95&mainannotation_id=29007.

2.2.2. Searching a Repository of Gene Expression Data

SymAtlas (symatlas.gnf.org/SymAtlas/) is a database of results from microarray experiments. It contains expression levels for genes of interest from a wide selection of tissue types. Results are collated from human and mouse tissues. Expression levels of genes with no previously known function are included in this database *(10)*.

To search SymAtlas, enter one or more accession numbers, gene names or gene ontology (GO) identifiers, separated by a space (*see* **Note 3**). The resulting histogram of gene expression levels is of most interest to the user, as they show which tissues or cell lines are expressing a gene and in which quantity (**Fig. 5**). The database actually contains expression information from a number



Symbols	RB1
Description	retinoblastoma 1 (including osteosarcoma)
Accessions	5925 (LocusLink)
Aliases	OSRC, RB, RB1
Phenotype	Bladder cancer, Osteosarcoma, Pinealoma with bilateral retinoblastoma, Retinoblastoma
Function	• androgen receptor binding (GO:0050681) • androgen receptor signaling pathway (GO:0030521) • negative regulation of protein kinase activity (GO:0006469) • negative regulation of transcription from RNA polymerase II promoter (GO:0000122) • positive regulation of transcription, DNA-dependent (GO:0045893) • protein binding (GO:0005515) • regulation of transcription, DNA-dependent (GO:0006355) • transcription (GO:0006350) • transcription coactivator activity (GO:0003713) • transcription factor activity (GO:0003700)
Protein Family	• Cyclin (IPR006670) • Retinoblastoma-associated protein, A-box (IPR002720) • Retinoblastoma-associated protein, B-box (IPR002719)
Genome Location	• Chr 13: 46.676-46.854 (+) (NCBI34) • Chr 13: 47.776-47.954 (+) (NCBI35) • Chr 13: 47.776-47.954 (+) (NCBI36)
Transcripts	• NM_000321.1 (RefSeq)
	• NP_000312.1 (RefSeq) • P06400 (SPTREMBL)

Fig. 5. Expression of the RB1 gene in different tissues, as documented in the SymAtlas database (symatlas.gnf.org/SymAtlas/). Here the expression levels are shown from the Human GeneAtlas GNF1H MAS5 dataset using the 203132_at reporter.

of experiments and sources, and the user may select what is appropriate by using the dataset selection panel on the top of the page. The gene expression is measured by using Affymetrix microarrays. The different gene expression levels may be viewed as separate histograms by selecting from the panel on the top of the page.

A brief browse through the *RB1* gene expression level in the Human GeneAtlas GNF1H, MAS5 dataset shows that it is under-expressed in normal bone marrow, as compared to other tissue types. *RB1* was under-expressed, albeit detectably, for both the mRNA reporters 203132_at (HG-U133A) and 211540_s_at (HG-U133A). Expression of the *RB1* gene was not detectable in retinoblastomas, osteosarcomas or soft tissue sarcomas ***(2)***. Therefore, the expression level of the *RB1* gene could be used to help determine the exact cause of these cancer types.

2.2.3. Understanding Subcellular Protein Localization

The human protein atlas project, above, is generating some information on the subcellular localization of proteins. For example, it is described in the annotation summary section for the RB1 protein, that RB1 is localized in the nucleus for almost all tissues, whether they are normal and malignant. However, there is no large-scale experiment which has been undertaken to date to accurately determine the precise sub-cellular localization of all human proteins. As an alternative, subcellular protein localization data can be obtained from studies on individual proteins. This is available in the Swiss-Prot database, and is systematized by the Gene Ontology.

The Swiss-Prot database can be queried for RB1_HUMAN. The GO terms in the Swiss-Prot database, under the cross-reference field, indicate that the protein is localized in the nucleus and the chromatin. The comments field of Swiss-Prot will sometimes also provide a description of the sub-cellular localization.

2.3. Where is the Protein Found in Biochemical Pathways, Cellular Reactions or the Reactome?

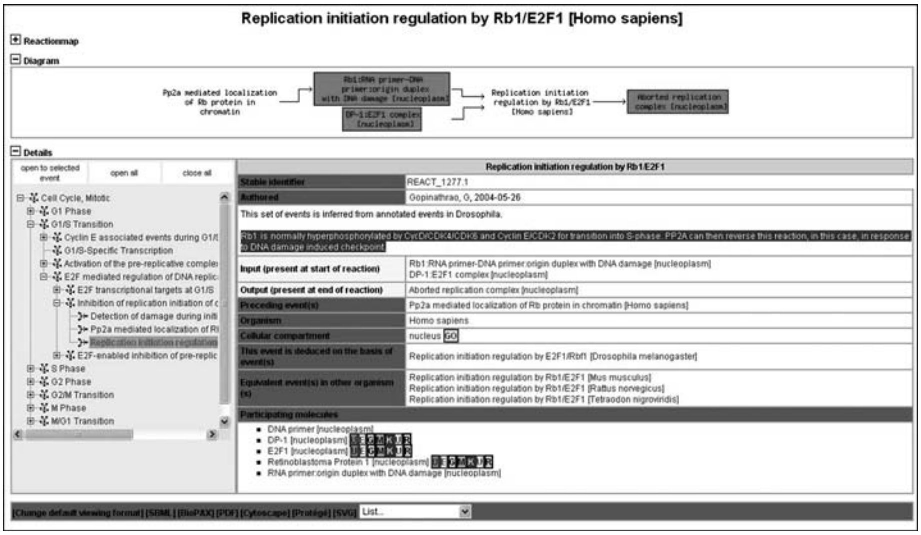
2.3.1. The Reactome Project

The reactome project (www.reactome.org) is a curated resource of pathways and reactions. It represents these pathways in graphical as well as tabular format. It draws on information from, and is hypertext linked to, other resources including KEGG (Kyoto Encyclopedia of Genes and Genomes), the Gene Ontology (GO) and the metabolite database (Chemical Entities of Biological Interest; ChEBI) ***(11,12)***.

The user may search for the context of a protein in the reactome by typing the gene name in the text box at the top of the front page. From the summary results page, click through to reaction section (*see* **Note 4**). The name RB1 can be used to search for the reactions described on that results page (*see* **Note 5**).

Searches of the reactome database pinpoint the exact reactions in which a protein participates. The reaction is described with a flow chart figure and is also described in text. The inputs of the reaction are described as well as the products: these include small molecules that may be substrates or products in enzyme-mediated reactions. The hypertext links to databases like KEGG and GO help the researcher to learn more about the reaction pathway(s) of interest.

For RB1, the reaction the protein is described as “Replication initiation regulation by Rb1/E2F1” (*see* **Fig. 6**) and the detail of the reaction is described: “Rb1 is normally hyperphosphorylated by CycD/CDK4/CDK6 and Cyclin


Reactionmap
Diagram
Pp2a mediated localization of Rb protein in chromatin
Rb1:RNA primer-DNA primer:origin duplex with DNA damage [nucleoplasm]
DP-1:E2F1 complex [nucleoplasm]
Replication initiation regulation by Rb1/E2F1 [Homo sapiens]
Aborted replication complex [nucleoplasm]
Details
open to selected event
open all
close all
Cell Cycle, Mitotic
G1 Phase
G1/S Transition
Cyclin E associated events during G1/
G1/S-Specific Transcription
Activation of the pre-replicative comple
E2F mediated regulation of DNA replic
E2F transcriptional targets at G1/S
Inhibition of replication initiation of c
Detection of damage during initi
Pp2a mediated localization of R
Replication initiation regulation
E2F-enabled inhibition of pre-replic
S Phase
G2 Phase
G2/M Transition
M Phase
M/G1 Transition
[Change default viewing format] [SBML] [BioPAX] [PDF] [Cytoscape] [Protégé] [SVG]
List...

Replication initiation regulation by Rb1/E2F1 [Homo sapiens]

Replication initiation regulation by Rb1/E2F1	
Stable identifier	REACT_1277.1
Authored	Gopinathrao, G, 2004-05-26

This set of events is inferred from annotated events in Drosophila.

Rb1 is normally hyperphosphorylated by CycD/CDK4/CDK6 and Cyclin E/CDK2 for transition into S-phase. PP2A can then reverse this reaction, in this case, in response to DNA damage induced checkpoint

Input (present at start of reaction)	Rb1:RNA primer-DNA primer:origin duplex with DNA damage [nucleoplasm] DP-1:E2F1 complex [nucleoplasm]
Output (present at end of reaction)	Aborted replication complex [nucleoplasm]
Preceding event(s)	Pp2a mediated localization of Rb protein in chromatin [Homo sapiens]
Organism	Homo sapiens
Cellular compartment	nucleus GO
This event is deduced on the basis of event(s)	Replication initiation regulation by E2F1/Rbf1 [Drosophila melanogaster]
Equivalent event(s) in other organism (s)	Replication initiation regulation by Rb1/E2F1 [Mus musculus] Replication initiation regulation by Rb1/E2F1 [Rattus norvegicus] Replication initiation regulation by Rb1/E2F1 [Tetraodon nigroviridis]

Participating molecules

- DNA primer [nucleoplasm]
- DP-1 [nucleoplasm] U E C M K I R
- E2F1 [nucleoplasm] U E C M K U R
- Retinoblastoma Protein 1 [nucleoplasm] U E C M K U R
- RNA primer:origin duplex with DNA damage [nucleoplasm]

Fig. 6. Summary results page from the reactome project for protein RB1. Note that this page immediately contextualizes the protein into a pathway, and gives details on the reaction inputs and outputs.

E/CDK2 for transition into S-phase. PP2A can then reverse this reaction, in this case, in response to DNA damage induced checkpoint."

2.3.2. The Gene Ontology

Gene ontology (GO) (www.geneontology.org) is a project aimed at providing controlled and consistent vocabulary for the description of gene and protein function, and a means to classify these into biologically meaningful categories. It is a global system and is applicable to all species. There are three categories in GO. They are cellular component, biological process and molecular function. For any gene or gene product, cellular component describes the part of a cell where the entity is found, for example, rough endoplasmic reticulum or proteasome. A biological process is a series of molecular functions or processes performed by assemblies of biomolecular entities, for example, signal transduction or pyrimidine metabolism. A molecular function describes an activity at a molecular level, for example, a catalytic activity, a type of binding, and more specifically, Toll receptor binding. The GO definitions are like a hierarchy, with the exception that a child GO term may have more than one parent. For instance, the term hexose biosynthesis has two parents: hexose metabolism and monosaccharide biosynthesis. If the child GO definition is annotated to a gene, the annotation automatically cascades to the parental GO term in a recursive manner ***(13)***.

Gene ontology is not a database of gene product names, or a database recording the attributes of sequences such as gene introns and exons. It does not describe protein tertiary structures or protein-protein interactions. Terms unrelated to the normal function of any gene, for example oncogenesis, are also not included. In addition, any descriptors that are above the level of cellular component, such as anatomical or histological features and cell types are not described. Other broad categories such as gene evolution and gene expression are similarly not addressed.

A useful web-based tool for browsing GO is quickGO (www.ebi.ac.uk/ego/index.html). A general method of searching is to type in the UniProt accession number of the gene (*see* **Note 6**). For the retinoblastoma protein, the UniProt accession number is P06400.

GO can provide insights into the molecular functions of a protein in one or more cellular processes. This helps the researcher find genes that are involved in similar molecular processes and functions, or are of the same cellular component. Genes with common GO terms also have a high chance of association through protein-protein interactions, or may be found in the same diseases or pathways. It is noteworthy that GO provides synonyms to terms of interest. This can provide researchers with appropriate keywords to effectively search other databases which do not use the GO vocabulary.

A search of the Swiss-Prot database for the RB1-associated GO terms reveals a set of classifications for RB1 cellular component, molecular function and biological process (*see* **Table 1**). Each of these classifications has an associated acyclic graph with the name of the GO term, all the GO terms' parents and the "ancestral" GO terms associated with them. These graphs can be accessed from quickGO (*see* **Fig. 7**), give a clear view of the hierarchy of the ontology for RB1 and illustrate the assigned function(s). In the case of RB1, this is "regulation of progression through the cell cycle." Note that it is common for one protein to map to more than one part of the gene ontology. In part, this is because proteins can be multifunctional, but also because a single protein can be involved in more than one process.

2.3.3. Searching BioCarta

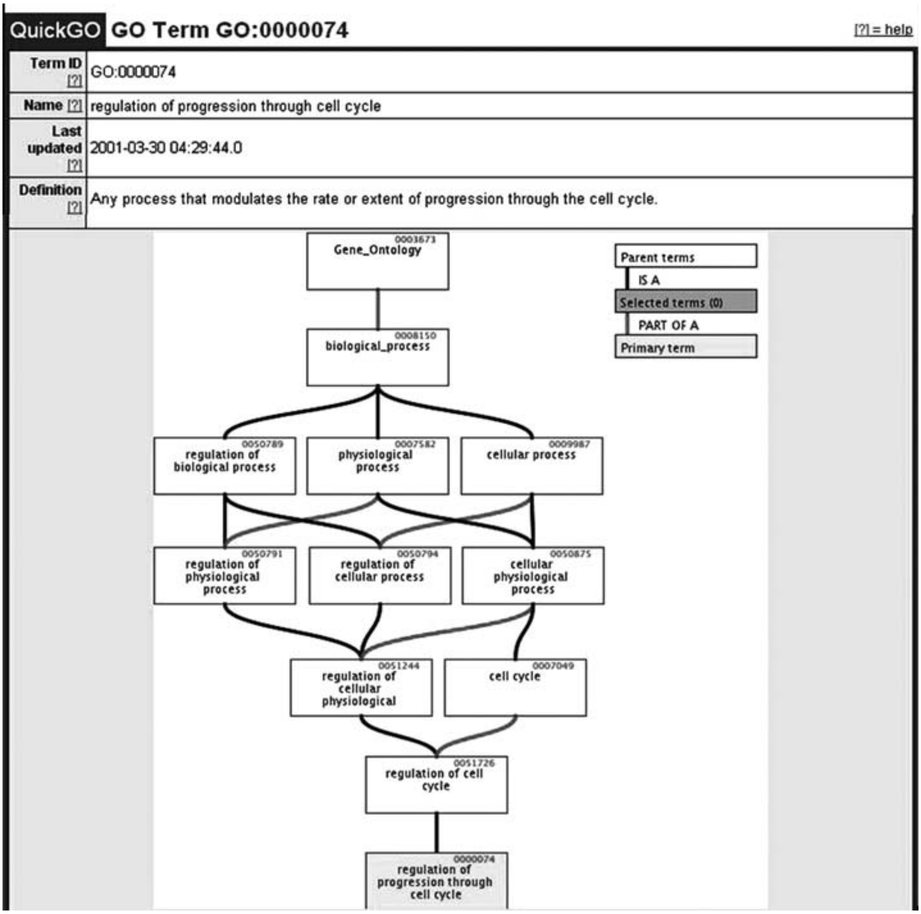
BioCarta is a curated database of biological pathways. Its particular strength is that it visualizes the pathway of interest and has legends that are easy to interpret. The icons for the protein provide links to other databases, such as

Table 1
Gene ontology (GO) terms assigned to the RB1 retinoblastoma-associated protein

GO Classification	Description
Cellular component	• chromatin • nucleus
Molecular function	• androgen receptor binding • protein binding • transcription coactivator activity • transcription factor activity
Biological process	• androgen receptor signaling pathway • cell cycle checkpoint • G1 phase • M phase • negative regulation of cell growth • negative regulation of protein kinase activity • negative regulation of transcription from RNA polymerase II promoter • positive regulation of transcription, DNA-dependent

OMIM and Swiss-Prot. It is different to the Gene Ontology as it does not seek to provide a global context for a particular gene or protein, instead focusing on the local environment and pathways in which a protein participates. For each pathway, a detailed textual description is given, as well as contact details for a "pathway expert".

To use BioCarta, the user first needs to access the main page at www.biocarta.com. Click on the "Pathways" tab at the top of the index page, and in the subsequent page, search by using the "gene name" text box under the section "search pathways by title". For example, the gene name "RB1" would be used for the retinoblastoma protein. This produces a list of pathways in which the protein of interest is involved. Choose the pathway of interest for investigation, by clicking on the pathway name. For the retinoblastoma tumour pathway, the pathway of most relevance is "RB Tumor Suppressor/Checkpoint Signaling in response to DNA damage." This is shown in **Fig. 8**.



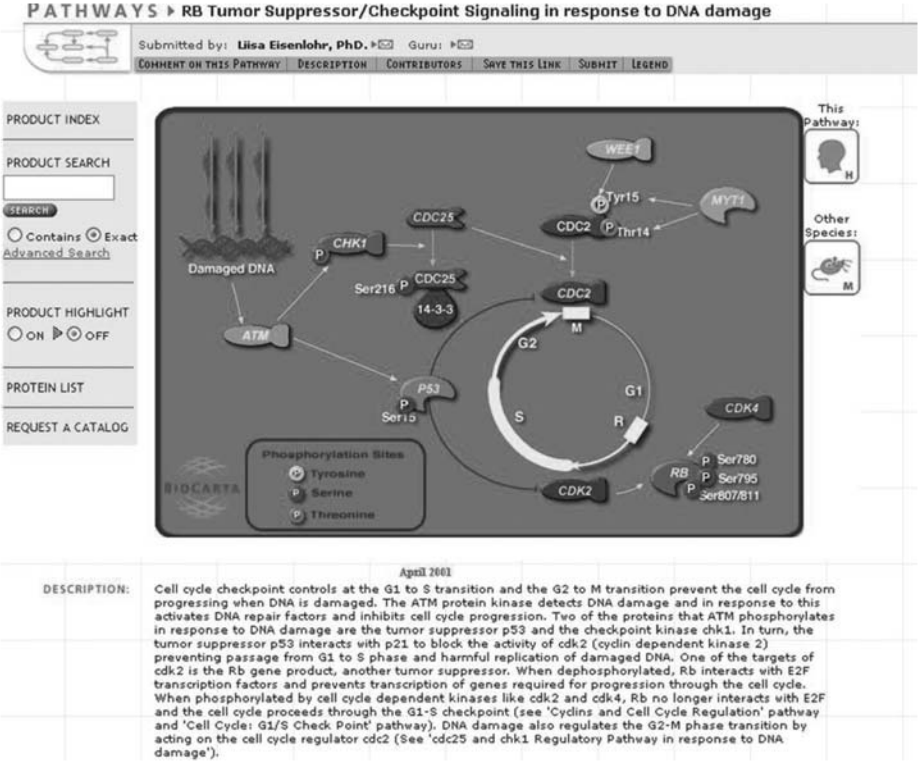
QuickGO GO Term GO:0000074 [?] = help

Term ID [?]	GO:0000074
Name [?]	regulation of progression through cell cycle
Last updated [?]	2001-03-30 04:29:44.0
Definition [?]	Any process that modulates the rate or extent of progression through the cell cycle.

Fig. 7. Acyclic graph showing how the RB1 protein maps onto the gene ontology (GO). This page can be accessed at www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0000074. Note that there are six levels to this part of the ontology (biological process). The lowest level of the ontology here is "regulation of progression through the cell cycle".

2.4. Are Any Protein-Protein Interactions Known?

Broadly speaking, there are two types of protein–protein interaction data. There are those from high throughput studies using techniques such as yeast two-hybrid or affinity purification of protein complexes and those that result from the study of individual proteins. For the latter, curation of literature can be used to generate a large dataset of interactions. For humans, there are two large,



PATHWAYS ▸ RB Tumor Suppressor/Checkpoint Signaling in response to DNA damage

Submitted by: Liisa Eisenlohr, PhD. Guru:

April 2001

DESCRIPTION: Cell cycle checkpoint controls at the G1 to S transition and the G2 to M transition prevent the cell cycle from progressing when DNA is damaged. The ATM protein kinase detects DNA damage and in response to this activates DNA repair factors and inhibits cell cycle progression. Two of the proteins that ATM phosphorylates in response to DNA damage are the tumor suppressor p53 and the checkpoint kinase chk1. In turn, the tumor suppressor p53 interacts with p21 to block the activity of cdk2 (cyclin dependent kinase 2) preventing passage from G1 to S phase and harmful replication of damaged DNA. One of the targets of cdk2 is the Rb gene product, another tumor suppressor. When dephosphorylated, Rb interacts with E2F transcription factors and prevents transcription of genes required for progression through the cell cycle. When phosphorylated by cell cycle dependent kinases like cdk2 and cdk4, Rb no longer interacts with E2F and the cell cycle proceeds through the G1-S checkpoint (see 'Cyclins and Cell Cycle Regulation' pathway and 'Cell Cycle: G1/S Check Point' pathway). DNA damage also regulates the G2-M phase transition by acting on the cell cycle regulator cdc2 (See 'cdc25 and chk1 Regulatory Pathway in response to DNA damage').

Fig. 8. The BioCarta illustration for the "RB Tumor Suppressor/Checkpoint Signalling in response to DNA damage" pathway, accessible from page www.biocarta.com/pathfiles/h_rbPathway.asp. Note that this pathway figure is accompanied by a detailed description.

high throughput studies to date (*see* **Note 7**) ***(14,15)***. Whilst extensive, these studies together represent less than a few percent of the human interactome. Accordingly, when searching for interaction partners of human proteins, it is necessary to use resources that consider interactions documented in small and large-scale studies. Databases such the HPRD (www.hprd.org) and IntAct contain such information. The IntAct database will be discussed here, because of the extensive cross-linking with other EBI resources ***(16)***.

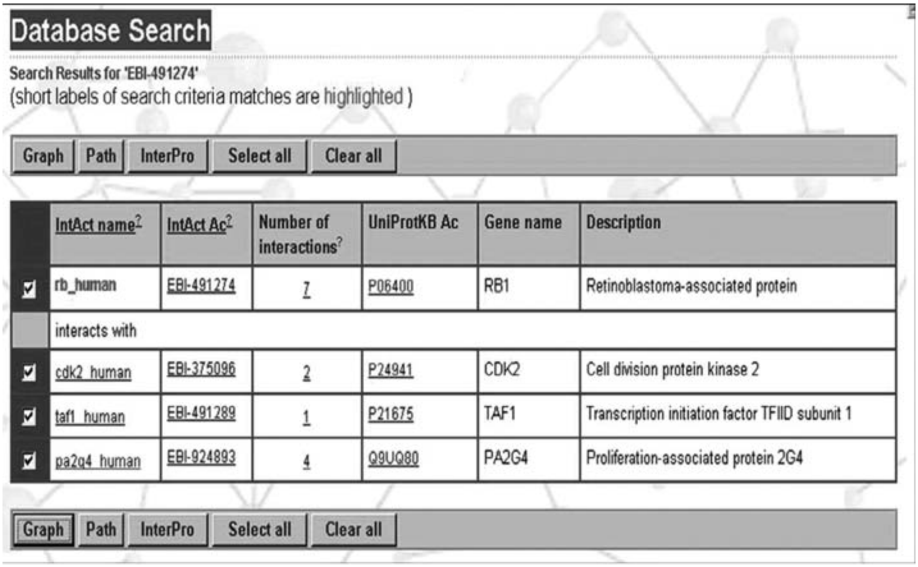
2.4.1. The IntAct Database of Protein–Protein Interactions

IntAct (www.ebi.ac.uk/intact/site/) is a database for protein-protein interaction data. The interaction data result from user submission or by the curation of published literature. IntAct allows the user to search for interaction partners for human proteins.

The user can search the IntAct database by entering a Swiss-Prot identification number in the search box on the front page. For retinoblastoma-associated protein, the RB_HUMAN identifier is used. The results of this type of search are the proteins which are known to directly interact with the protein of interest. In the case of RB_HUMAN, these are proteins Cdk2, Taf1, and Pa2g4 (*see* **Fig. 9**).

It is usually of interest to visualize the interactions that a protein participates in and how this fits into a local or global interaction network. IntAct allows interactions to be visualized as scale-free graphs with the Hierarch viewer. Proteins of interest are selected from the list of interacting proteins (**Fig. 9**) and the graph button then selected (*see* **Note 8**). The resulting scale-free graphs display the proteins as nodes (protein names) and the interactions between proteins as edges (lines). The protein of interest appears in the center of the graph. **Figure 10** shows the local interaction network for the retinoblastoma-associated protein.

The Hierarch viewer, as described above, also provides additional contextual information for a protein and its interactors. A list of GO terms and protein

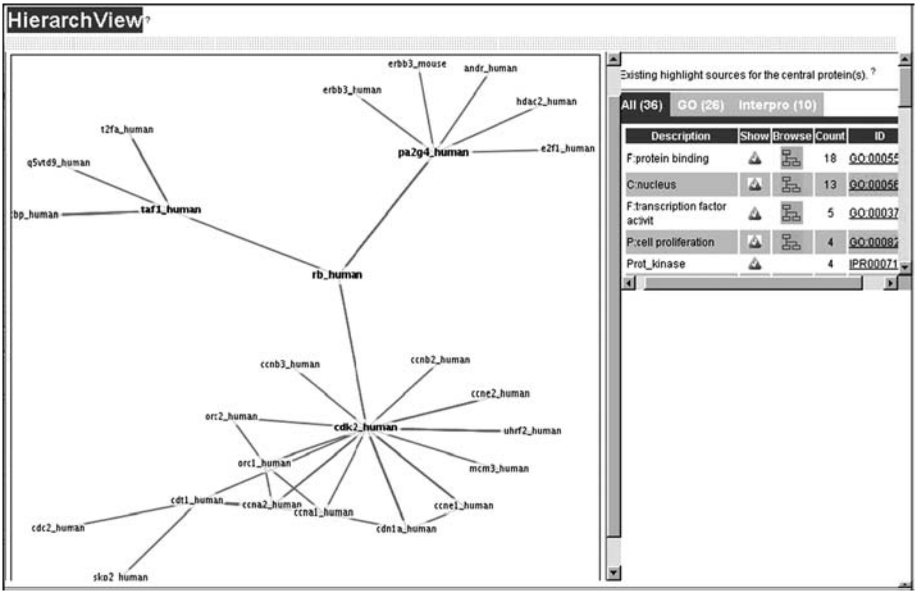


Database Search

Search Results for 'EBI-491274'
(short labels of search criteria matches are highlighted)

	IntAct name$^{?}$	IntAct Ac$^{?}$	Number of interactions$^{?}$	UniProtKB Ac	Gene name	Description
☑	rb_human	EBI-491274	7	P06400	RB1	Retinoblastoma-associated protein
	interacts with					
☑	cdk2_human	EBI-375096	2	P24941	CDK2	Cell division protein kinase 2
☑	taf1_human	EBI-491289	1	P21675	TAF1	Transcription initiation factor TFIID subunit 1
☑	pa2g4_human	EBI-924893	4	Q9UQ80	PA2G4	Proliferation-associated protein 2G4

Fig. 9. Search results from the IntAct database for the retinoblastoma-associated protein. This shows that the protein directly interacts with three other proteins.



Existing highlight sources for the central protein(s). ?

Description	Show	Browse	Count	ID
F:protein binding	▲	▨	18	GO:00055
C:nucleus	▲	▨	13	GO:00056
F:transcription factor activit	▲	▨	5	GO:00037
P:cell proliferation	▲	▨	4	GO:00082
Prot_kinase	▲		4	IPR00071

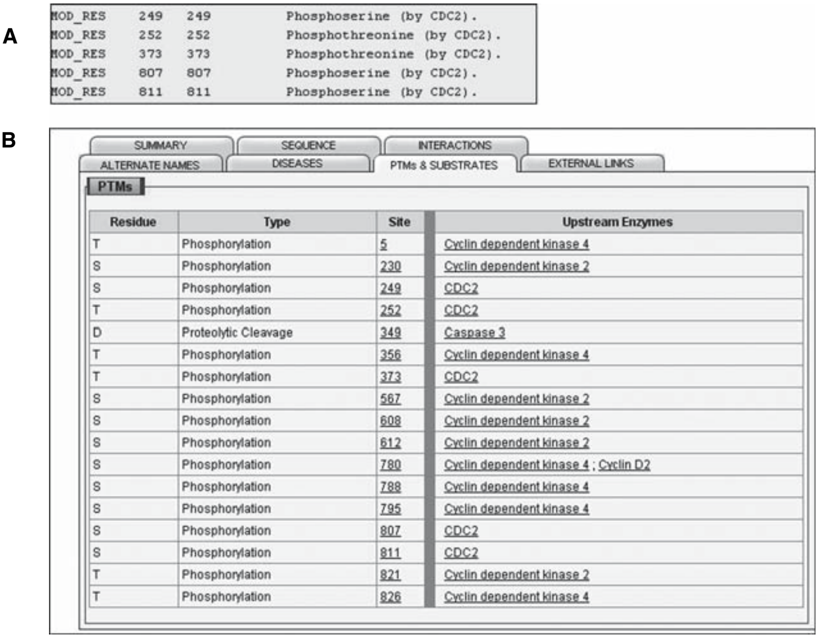
Fig. 10. Contextualization of the retinoblastoma-associated protein in the local protein-protein interaction network. The RB_HUMAN protein is at the center of the view, and other proteins up to two interactions away from it are shown. The right-hand side of the viewer provides access to links to the gene ontology and domain-based information.

functional domains from InterPro (*see* **Note 9**) *(17)* is given on the right hand side of the viewer web page (*see* **Note 10**). Neighboring proteins are likely to be involved in similar functions to the protein of interest and may contribute to the molecular basis for the development of the disease. The importance of a protein may be related to the number of interactions in the protein-protein interaction network. Highly connected proteins are thought to be more highly associated with disease.

2.5. What Types of Post-Translational Modifications does the Protein Carry?

2.5.1. The Swiss-Prot Database

As a final consideration, it can be of interest to understand the post-translational modifications that are known to be carried by a protein. This can provide clues to protein localization and function. The Swiss-Prot database is an excellent source of post-translational modification information, where all modifications are collated from the literature. To find post-translational modifications in Swiss-Prot, the user needs to refer to relevant MOD_RES annotations in the feature table of each database entry. Various modifications are described therein, including fatty acids, glycosylation and reversible modifications, including phosphorylation, methylation, and acetylation. Disulfide bridges of proteins are also documented in this part of the database. For the retinoblastoma-associated protein, Swiss-Prot documents phosphorylation at amino acid positions 249, 252, 373, 807, and 811. It also notes that this phosphorylation is mediated by the kinase CDC2 (*see* **Note 11**). **Figure 11A** shows the relevant portion of the feature table for the retinoblastoma-associated protein.



```
MOD_RES     249    249          Phosphoserine (by CDC2).
MOD_RES     252    252          Phosphothreonine (by CDC2).
MOD_RES     373    373          Phosphothreonine (by CDC2).
MOD_RES     807    807          Phosphoserine (by CDC2).
MOD_RES     811    811          Phosphoserine (by CDC2).
```

Residue	Type	Site	Upstream Enzymes
T	Phosphorylation	5	Cyclin dependent kinase 4
S	Phosphorylation	230	Cyclin dependent kinase 2
S	Phosphorylation	249	CDC2
T	Phosphorylation	252	CDC2
D	Proteolytic Cleavage	349	Caspase 3
T	Phosphorylation	356	Cyclin dependent kinase 4
T	Phosphorylation	373	CDC2
S	Phosphorylation	567	Cyclin dependent kinase 2
S	Phosphorylation	608	Cyclin dependent kinase 2
S	Phosphorylation	612	Cyclin dependent kinase 2
S	Phosphorylation	780	Cyclin dependent kinase 4 ; Cyclin D2
S	Phosphorylation	788	Cyclin dependent kinase 4
S	Phosphorylation	795	Cyclin dependent kinase 4
S	Phosphorylation	807	CDC2
S	Phosphorylation	811	CDC2
T	Phosphorylation	821	Cyclin dependent kinase 2
T	Phosphorylation	826	Cyclin dependent kinase 4

Fig. 11. Post-translational modifications documented for the retinoblastoma-associated protein: **(A)** Portion of the Swiss-Prot feature table for entry RB_HUMAN, detailing which amino acids are phosphorylated. **(B)** Similar entry from the HPRD database. *See* text for details.

2.5.2. The Human Protein Reference Database

The Human Protein Reference Database (HPRD) (www.hprd.org) is a database which centralizes protein annotation. It contains information such as functional domains, co- and post-translational modifications, protein-protein interactions and associations with mutations and disease. The information is curated by experts that mine literature references and published data. These information are accessible via a web-based interface ***(18)***.

Unlike the Swiss-Prot database, the HPRD database only provides annotation and sequence information for human proteins. As a result of this focus, HPRD may provide more post-translational modification data than Swiss-Prot for human proteins. It also provides detailed information on the enzyme which catalyses the addition of the post-translational modification (PTM) on a protein, and useful links to the literature reference(s) that discovered the modification. This latter feature is not available in the Swiss-Prot database. Hypertext link to literature references for information such as tissue distribution, subcellular localization and disease association of proteins are also provided ***(18)***.

To access modification information in HPRD, the user first needs to access the query interface via the databases front page (www.hprd.org). This is done by selecting the query button on the top left-hand corner of the page. The user may then search HPRD for the protein of interest by using a Swiss-Prot accession number, in this case P06400 for the retinoblastoma-associated protein (*see* **Note 12**). At the results page, click on the tab for "PTMs and Substrates."

A small cartoon in the PTM annotation page illustrates approximately where each post-translational modification is found on the protein. The upstream enzymes thought to be responsible for the addition of the PTM are noted for each modified amino acid. Each modification is linked to a literature resource, which can be seen by clicking on the amino acid position number in the table. The RB1 protein is phosphorylated at 16 positions in the HPRD database (**Fig. 11B**), as compared to five positions as annotated in Swiss-Prot. It is mainly phosphorylated by proteins that regulate cell division cycles: cell division cycle 2 (CDC2 or CDK1), Cyclin D2, and two cyclin dependent kinases CDK2 and CDK4.

2.5.3. Predicted Post-Translational Modifications

Databases of predicted post-translational modifications may be relevant and of interest to the researcher. Predictions in these databases must, however, be used with caution as the predictions may be of low quality. They may provide false predictions and also neglect real modification sites. Databases containing predicted modifications include the dbPTM (http://dbptm.mbc.nctu.edu.tw/index.html), which contains predicted modifications for many proteins, including the retinoblastoma-associated protein ***(19)***.

3. Notes

1. Human Protein Atlas—Normal tissue: It is often difficult to obtain normal human tissues, since they are derived from surgical material. Therefore, *normal* is defined in this context as close to normal and samples would include alterations due to inflammation, degeneration and tissue remodeling.
2. Human Protein Atlas—How to interpret reliability of results? A number of colored circles are found beside the name of the tissue type, whether they are normal or malignant tissues. Each colored circle represents the specific tissue type from one individual. A different color code is assigned to annotate the intensity and abundance of immunoreactivity (red = strong, orange = moderate, yellow = weak, white = no staining, and black = missing tissue). The circle is divided evenly such that each section represents replicate samples from the same tissue type.
3. Usage of SymAtlas—Wildcard characters ? and *, which represent one and any number of characters correspondingly, may be used. The search results list appears on the side panel of the webpage. Click on the little picture icon next to the appropriate gene name, under the *Homo sapiens* section. This will link to the gene expression chart. Clicking on the gene name will lead to a list of annotations and hyperlinks to other databases. The gene expression histogram may be accessed by selecting a dataset under the functional data section.
4. Usage of Reactome—Upon searching using the gene name, the results page will be shown. The matches would be classified into categories, for example: physical entity, reference entity, summation, reaction coordinates and reaction.
5. Reactome's page for RB1—The search for RB1 should get to this page: http://www.reactome.org/cgi-bin/eventbrowser?DB=gk_current&ID=113643& ***(20)***
6. QuickGO search result list—The search result will show every match to the query within the categories of biological process, molecular function and cellular component. Each GO term is attached to an identification number in the format GO: 7 digit number. An example GO ID number for the RB1 protein is: "regulation of progression through cell cycle, GO:0000074."
7. Protein-protein interactions—RB1 protein is part of the Rual et al. yeast-two-hybrid study on human proteins ***(14)***. However, it is not present in Stelzl et al. human yeast two-hybrid study ***(15)***.
8. IntAct search result list—You may need to select a number of interactor proteins to see the context of a protein in its interaction network. The "Select All" or "Clear All" button may be used to select and deselect the whole list of proteins**.** The "Path" button will show the minimal number of protein-protein interactions, or the minimally connecting network, for the selected set of proteins.

9. Interpreting IntAct results—A protein domain is a polypeptide chain that can fold autonomously into a structural unit. Some domains have a common evolutionary origin and molecular function ***(21)***.
10. Interpreting IntAct results—The list of GO terms can assist in understanding the biological function of the protein-protein interaction sub-network. The "show" button beside each GO and InterPro entry on the right-hand side of the web-page allows the user to highlight proteins with that annotation in the graph. The count indicates the number of proteins in the graph to which the GO or InterPro entry applies. The user can click the hypertext links to browse the details of the relevant GO terms or access biological information about protein domains of interest.
11. Protein phosphorylation**—**CDC2 is shown to interact indirectly with RB1 through CDK2 in Figure 10. This may be due to a missing interaction not documented in IntAct. Another explanation would be that phosphorylation of RB1 by CDC2 is dependent on CDK2, such that phosphorylation would only occur if the three proteins are in a complex.
12. Using the HPRD database**—**There are many other methods of searching the database, for example, searching by PTMs, molecular class and cellular component. The user may browse the database by clicking the "Browse" button on the top left-hand corner of the web page, and accessing the browsing interface. This would give the list of all possible searches available for each category, for instance, functional domains, PTMs and cellular sub-localization.

References

1. Fung, Y. K., Murphree, A. L., T'Ang, A., Qian, J., Hinrichs, S. H., and Benedict, W. F. (1987) Structural evidence for the authenticity of the human retinoblastoma gene. *Science* **236,** 1657–1661.
2. Weichselbaum, R. R., Beckett, M., and Diamond, A. (1988) Some retinoblastomas, osteosarcomas, and soft tissue sarcomas may share a common etiology. *Proc. Natl. Acad. Sci. U S A* **85,** 2106–2109.
3. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159.
4. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191.
5. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517.
6. Buchkovich, K., Duffy, L. A., and Harlow, E. (1989) The retinoblastoma protein is phosphorylated during specific phases of the cell cycle. *Cell* **58**, 1097–1105.

7. Chen, P. L., Scully, P., Shew, J. Y., Wang, J. Y., and Lee, W. H. (1989) Phosphorylation of the retinoblastoma gene product is modulated during the cell cycle and cellular differentiation. *Cell* **58**, 1193–1198.
8. DeCaprio, J. A., Ludlow, J. W., Lynch, D., Furukawa, Y., Griffin, J., Piwnica-Worms, H., et al. (1989) The product of the retinoblastoma susceptibility gene has properties of a cell cycle regulatory element. *Cell* **58**, 1085–1095.
9. Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A., Amini, B., Andersen, E., et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **4**, 1920–1932.
10. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., et al. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4465–4470.
11. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., et al. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432.
12. Joshi-Tope, G., Vastrik, I., Gopinath, G. R., Matthews, L., Schmidt, E., Gillespie, M., et al. (2003) The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 237–243.
13. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
14. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178.
15. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968.
16. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455.
17. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201–D205.
18. Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371.
19. Lee, T. Y., Huang, H. D., Hung, J. H., Huang, H. Y., Yang, Y. S., and Wang, T. H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.* **34**, D622–D627.

20. Gopinathrao, G. (2004) Replication initiation regulation by Rb1/E2F1 [Homo sapiens]. Reactome project. Viewed 20 September 2006 http://www.reactome.org/cgi-bin/eventbrowser?DB=gk_current&ID=113643&.
21. Mount, D. M. (2001) *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, New York, NY.

17

Web-Based Resources for Clinical Bioinformatics

Anthony M. Joshua and Paul C. Boutros

Summary

In the post-Human Genome Project era, awareness of the resources available through the internet is essential to both molecular biologists and clinicians. An overview of the main databases and analytical tools described in this chapter is important to understand the principles upon which hypotheses are generated, experiments are based and conclusions reached. Similarly, an introduction to the terminology of these resources often facilitates their use and adoption into practice. This chapter covers database resources such as NCBI/ Entrez, Ensembl and UCSC as well as analytical tools for sequence alignment, promoter analysis and molecular interactions.

Key Words: internet, database, bioinformatics, web servers, biological sequences, genome-browser, sequence-alignments.

Abbreviations: BLAST – Basic Local Alignment Search Tool; DDBJ –DNA Data Bank of Japan; EMBL – European Molecular Biology Laboratory; EST – Expressed Sequence Tag; NCBI – National Centre for Biotechnology Information; UCSC – University of California, Santa Cruz

1. Introduction

It is an oft-repeated truism that the last three decades has seen an unparalleled expansion in biological data gathering. Managing and analyzing this information was impossible before the advent of the internet. Putting the collated and organized information of the previous decades at the fingertips of researchers has had a significant impact on both the pace and depth of our understanding of biological processes both at the bench and the bedside.

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

The history of biological databases began in the early 1960s, with the publication of the *Atlas of Protein Sequence and Structure* by Margaret Dayhoff and colleagues ***(1)***. By 1972 the contents of this volume were too large to be printed and were distributed electronically on magnetic tape. Following the arrival of DNA sequence databases in the early 1980s, the formation of the International Nucleotide Sequences Database Collaboration (GenBank, The National Institutes of Health genetic sequence database; EMBL, European Molecular Biology Laboratory; and DDBJ, DNA Data Bank of Japan) and agreement upon a common sequence-annotation format allowed for the simultaneous synchronization of these three databases.

In parallel to the nucleotide databases, protein databases were formed, initially under the guidance of Amos Bairoch at the University of Geneva. This eventually led to the formation of Swiss-Prot in 1986 (*see* **Chapter 16** for more discussion on Swiss-Prot).

Today, both the original primary (archival) and evolving secondary (curated) databases with their associated analytical tools are indispensable in the modern laboratory. Learning how to use the web-based databases is a hands-on experience. While we will try to emphasize issues of overall relevance, the purpose of this chapter is to give the reader an overview of the main sequence databases and analytical tools on the web. A detailed description or exhaustive list is beyond the scope of this chapter but exists in well-known bioinformatics textbooks ***(2)*** and yearly supplements to *Nucleic Acid Research* ***(3)***, respectively.

The reader is encouraged to review the websites in **Table 1** and attempt worked examples to form a greater appreciation of the various database's utility in their field of research. A brief overview of the databases in **Table 1** follows. If a particular area of interest is not represented at all or in sparse detail, web pages such as the ExPASy Life Science Directory offer a good overview of database resources available.

1.1. Sequence Databases—Introduction and Principles

The sequence databases of the International Nucleotide Sequences Database Collaboration are organized into divisions, which traditionally were based on taxonomies. Whilst the majority of these are consistent across all three databases, i.e., PRI for primate and PLN for plant, there are a few notable exceptions. For example, ORG (organelle) and PRO (prokaryotic) are only present in EMBL, while the HUM (human) division in DDBJ and EMBL are included in the PRI (primate) division in GenBank.

Table 1
Useful web sites for databases

URL	Name
www.ncbi.nih.gov	NCBI
www.ensembl.org	Ensembl
genome.ucsc.edu	UCSC Genome Browser
www.genecards.org	GeneCards
http://bioinfo2.weizmann.ac.il/geneloc/index.shtml	GeneLoc
www.genelynx.org	GeneLynx
eugenes.org	EuGenes
ca.expasy.org/links.html	ExPASy LifeScience Directory
www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene	*Entrez* Gene

Fortunately, the functional divisions are more consistent across the databases. These include ESTs (expressed sequence tags; short 300–500-bp single reads from mRNAs, which may not all be coding sequences), STS (sequence tagged sites; short 200–500-bp unique sequences that identify the combination of primer pairs used in a PCR assay, generating a reagent that is located to a single position in the genome), GSS (genome survey sequences which are similar to the EST division, except the sequences are genomic in origin), HTG (high throughput genomic sequences which are unfinished DNA genomic sequences generated by high-throughput sequencing centres), PAT (patent sequences), CON (constructed records of chromosomes, genomes and other long DNA sequences which include instructions on how to assemble pieces present in other divisions into a larger piece).

For cross-platform portability, the sequence information in all databases is stored in "flat-files"—plain text files—rather than proprietary database formats. These flat-files can be separated into three main parts: the header (contains the information or descriptors that apply to the record), the features (annotations of the record), and the nucleotide sequence itself. Important features of the header are the identification tags of the sequence. The accession number of a sequence is the number that is always linked to a particular record. The number is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits, e.g., U12345, or two letters followed by six digits e.g., AF123456. The version tag is made of the accession number of the database record followed by a dot and a version number (and is therefore sometimes referred to as "accession.version"). It is now the preferred method to refer to a particular sequence across all three databases. The GI (GenInfo) number is a series of digits that are assigned consecutively to each sequence record, and is processed by NCBI only. The GI number has no relationship to the Accession number of the sequence record.

All of the major sequence databases contain core features such as detailed gene information, a genomic-mapping interface, and associated analytical tools. However, it is important to remember that each of the databases and browsers discussed below presents a particular view of the human genome. In some cases it may be worthwhile to look at the same region with different web interfaces because there are some tools unique to each site. For example, GenBank does not allow researchers to display their own data in the context of a genome assembly, but it does provide other mapping resources such as the Mitelman chromosomal aberration database and the Stanford human hybrid cells map. Additionally, because the three sites discussed below use different methods to align mRNAs and ESTs to the genome, as well as different gene prediction algorithms, the positions and characteristics of the genes often vary.

1.2. Database Browsers

1.2.1. NCBI (National Center for Biotechnology Information)

The key to accessing information through the NCBI website is the *Entrez* system (**Fig. 1**). The existence of relationships between a gene and its various annotations—such as alternative transcripts, genetic locus, protein structure, mutations or pathology—lead to the formation of an integrated search interface. The *Entrez* system uses an easy to absorb interface and exploits simple searches with Boolean operators (such as *and* or *not*) for queries. The *Entrez* search result also briefly describes the nature of each database should the user require more information ***(4)***.

An alternative way to conceptualize collating genetic data is to organize it through genetic loci. NCBI's "*Entrez* Gene" ***(5)*** provides this facility allowing an overview of key connections in the map, sequence, expression, structure, function, citation, and homology data.

Building up from GenBank, which is a very redundant database of unordered sequences, an important advance to eliminate the redundancy in this led to the introduction of the RefSeq collection ***(6)***. In RefSeq, each biological entity (DNA, mRNA or protein) is represented once and only once (it is a non-redundant dataset). Their distinct accession numbers distinguish RefSeq entries from other entries at NCBI; "NT_" indicating genomic contigs, "NM_" indicating mRNA and "NP_" representing protein sequences. There are also computational predictions, which start with either "XM_" or "XP_" for model mRNAs and proteins respectively.

An important interface for scientists is often visualizing the genomic context of a gene in question. NCBI's viewing interface is called "MapViewer" and offers a tight integration with other NCBI resources allowing users to examine features in depth without leaving the site.

1.2.2. Ensembl

Certainly the most visually attractive of all the web-based interfaces, Ensembl is rather intuitive to use (**Fig. 2a**). There is a thorough annotation pipeline and a strong emphasis at EBI (European Bioinformatics Institute) on producing genome annotations in a timely manner with a high specificity ***(7)*** , in order not to over predict genome features.

By default, Ensembl displays four types of transcripts: (1) Ensembl transcripts (predicted by Ensemble), (2) Vega transcripts (from the Vertebrate

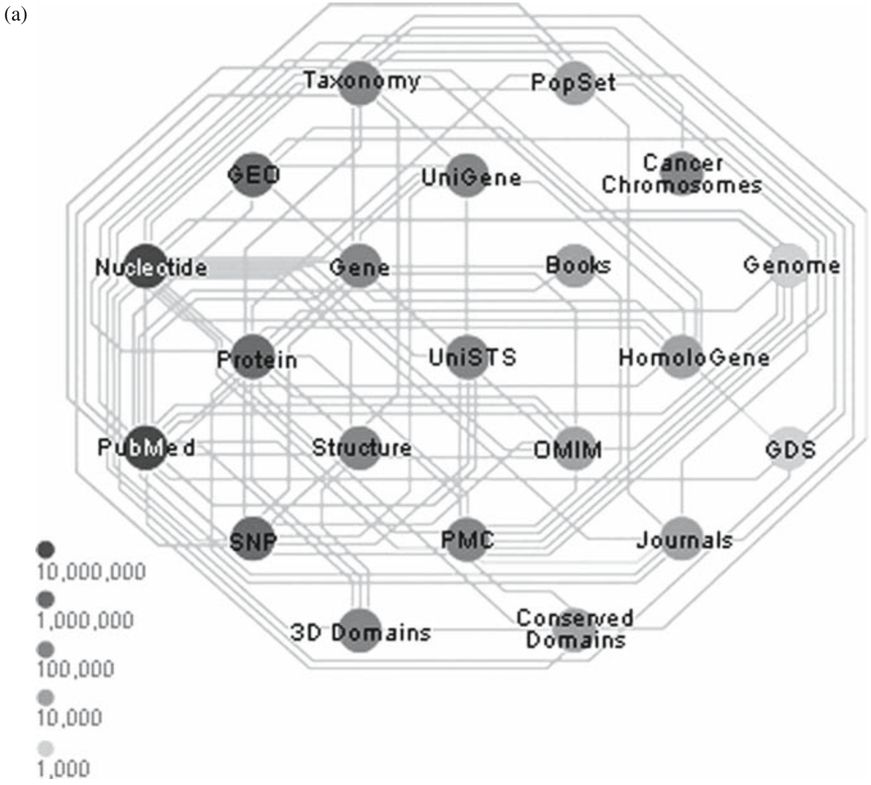
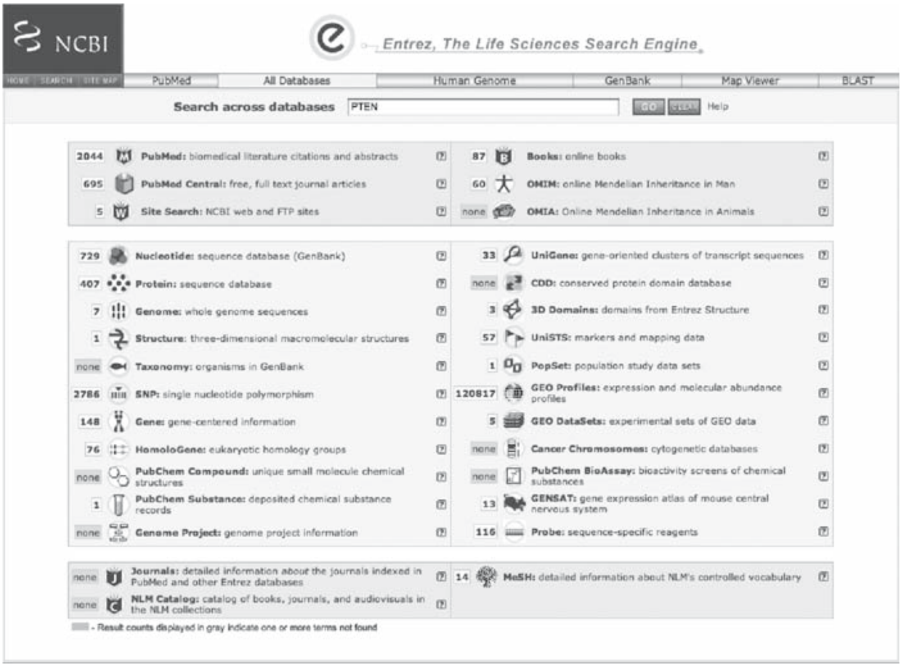


Fig. 1. *(Continued)*

(b)



Search across databases PTEN

2044	**PubMed:** biomedical literature citations and abstracts	87	**Books:** online books
695	**PubMed Central:** free, full text journal articles	60	**OMIM:** online Mendelian Inheritance in Man
5	**Site Search:** NCBI web and FTP sites	none	**OMIA:** Online Mendelian Inheritance in Animals
729	**Nucleotide:** sequence database (GenBank)	33	**UniGene:** gene-oriented clusters of transcript sequences
407	**Protein:** sequence database	none	**CDD:** conserved protein domain database
7	**Genome:** whole genome sequences	3	**3D Domains:** domains from Entrez Structure
1	**Structure:** three-dimensional macromolecular structures	57	**UniSTS:** markers and mapping data
none	**Taxonomy:** organisms in GenBank	1	**PopSet:** population study data sets
2786	**SNP:** single nucleotide polymorphism	120817	**GEO Profiles:** expression and molecular abundance profiles
148	**Gene:** gene-centered information	5	**GEO DataSets:** experimental sets of GEO data
76	**HomoloGene:** eukaryotic homology groups	none	**Cancer Chromosomes:** cytogenetic databases
none	**PubChem Compound:** unique small molecule chemical structures	none	**PubChem BioAssay:** bioactivity screens of chemical substances
1	**PubChem Substance:** deposited chemical substance records	13	**GENSAT:** gene expression atlas of mouse central nervous system
none	**Genome Project:** genome project information	116	**Probe:** sequence-specific reagents
none	**Journals:** detailed information about the journals indexed in PubMed and other Entrez databases	14	**MeSH:** detailed information about NLM's controlled vocabulary
none	**NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections		

▨ - Result counts displayed in gray indicate one or more terms not found

Fig. 1. **a:** An overview of the databases linked together through the *Entrez* system. **b:** An example of a search for the gene *PTEN*.

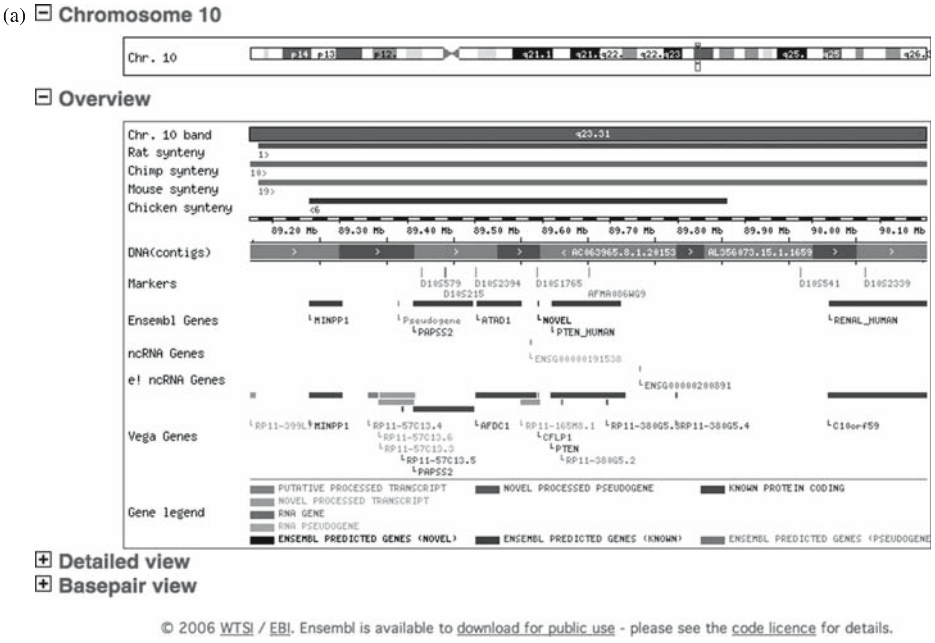
Genome Annotation (VEGA) consortium (a manually curated database)), (3) EST transcripts (based on EST evidence), and (4) GenScan Genes (gene predictions based on the gene prediction program GENSCAN). Along with the UCSC browser (described below) Ensembl allows for DAS (distributed annotation system) integration, which enables a researcher to gather genome annotation information from multiple distant web sites and display them in a single view. DAS is also currently used by the Ensembl browser to cover proteomic annotations and annotation of non-positional features.

A nice feature of the Ensembl site is the BioMart data-mining feature that allows the export of answers to complex, potentially genome wide queries just by filling in three pages on the web site (START—to define the dataset; FILTER—to apply combinations of various properties and OUTPUT—to choose properties of the filtered data set to be exported).

There are two features that make Ensembl particularly useful for high-level analysis. First, the software and underlying databases are available for download, enabling a scientist to run searches and analysis on their own computer, clearly facilitating access issues. Second, there are both Java and Perl “Application Programming Interfaces” or “APIs” which to external users may be useful to automate the extraction of particular data, to customize Ensembl to fulfill a particular purpose.

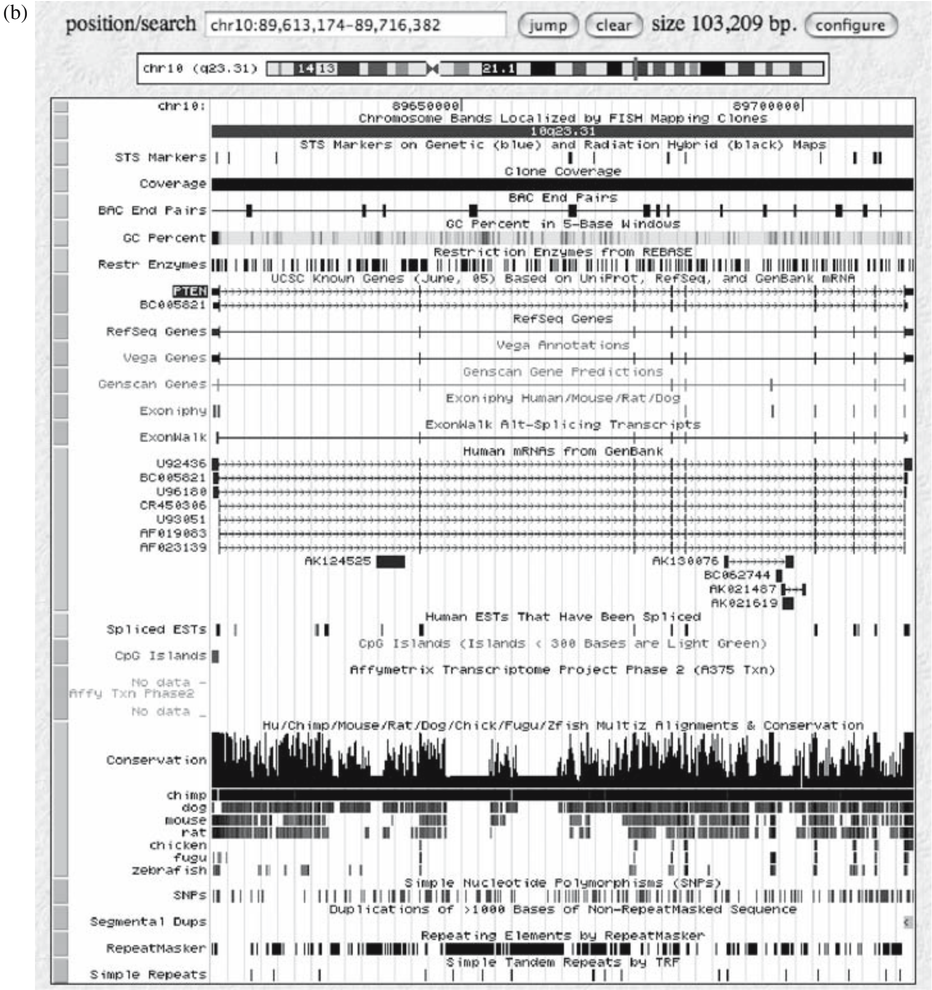
1.2.3. UCSC (University of California, Santa Cruz)

The UCSC genome browser *(8)* is based on annotating the genomic backbone with various levels of information in layers called “tracks.” Each track represents a different feature such as known genes, evolutionary conservation or single nucleotide polymorphisms (SNPs) and can be submitted by outside researchers (**Fig. 2b**). Additionally, the “Table browser” feature allows all the data to be downloaded in text form and offers the possibility of combining unions or intersections of tracks, which is a very useful tool. These features offer a powerful way to display rapidly a wide range of genomic data. For example, a particularly useful aspect of the UCSC interface is the ability to assess graphically multiple EST alignments or likely real transcripts using the



© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

Fig. 2. *(Continued)*



position/search chr10:89,613,174-89,716,382 (jump) (clear) size 103,209 bp. (configure)

(b)
chr10 (q23.31)
14
13
21.1
chr10:
89650000
89700000
Chromosome Bands Localized by FISH Mapping Clones
10q23.31
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Clone Coverage
Coverage
BAC End Pairs
BAC End Pairs
GC Percent in 5-Base Windows
GC Percent
Restriction Enzymes from REBASE
Restr Enzymes
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
PTEN
BC005821
RefSeq Genes
RefSeq Genes
Vega Annotations
Vega Genes
Genscan Gene Predictions
Genscan Genes
Exoniphy Human/Mouse/Rat/Dog
Exoniphy
ExonWalk Alt-Splicing Transcripts
ExonWalk
Human mRNAs from GenBank
U92436
BC005821
U96180
CR450306
U93051
AF019083
AF023139
AK124525
AK130076
BC062744
AK021487
AK021619
Human ESTs That Have Been Spliced
Spliced ESTs
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
Affymetrix Transcriptome Project Phase 2 (A375 Txn)
No data -
Affy Txn Phase2
No data _
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
chimp
dog
mouse
rat
chicken
fugu
zebrafish
Simple Nucleotide Polymorphisms (SNPs)
SNPs
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
Simple Repeats

Fig. 2. **a:** A snapshot of the Ensembl browser. **b:** A snapshot of the UCSC genome browser interface demonstrating the "tracks" of information under the gene *PTEN* on Chromosome 21q22.2.

"Alt-Splicing" track which summarizes splicing information by showing only exons and splice junctions that have an orthologous exons or splice junctions in the mouse, or that are present three or more times. Other aspects that researchers may find of use are the FTP site where sequences and their annotations can be downloaded for desktop analysis or the excellent online training available to explore its many features ***(9)***.

Finally, a minor but important point to understand is that the naming conventions for the human gene assemblies differs between UCSC and NCBI so for example, build 36.1 happens to be the 36th assembly of the genome entered at NCBI but corresponds to the March 2006 or hg18 UCSC assembly.

1.3. Gene Integration Resources

Similar to the concept behind "*Entrez* Gene" there are now a number of web-based resources that provide hyperlinked summaries of gene-related information so that a researcher can initially get a quick overview before choosing to go into greater depth on a particular issue. Only four are mentioned here but there are many others available. (1) *GeneCards* ***(10)*** provides a succinct summary of various gene attributes organized in an ease to read manner. (2) *GeneLoc* ***(11)*** provides useful summaries of a gene's genomic geography. (3) *GyneLynx* is another useful resource ***(12)***. There are two types of queries possible; a standard search term and a DNA sequence search that can accept raw sequences, accession IDs, or sequence files that can be uploaded. The breadth of information presented and collated on one page makes it a useful starting point for gathering information about a gene. (4) *euGenes* ***(13)*** (Genomic Information for Eurkaryotic Organisms) has a multi-species search interface and also provides integration of gene ontology (GO) selection criteria with other gene attributes.

2. Analytical Tools

Bioinformatics researchers have developed a wide range of tools for analyzing biological sequence data. Often different groups will develop very similar tools—just as broadly comparable genome browsers have been developed. Because these tools are so similar, often it is less important which tool is selected and more important what "parameters" are used for the analysis. These parameters can control many different aspects of the analysis, such as sensitivity/specificity trade-offs, computational efficiency, and output characteristics. In this section, we overview both the major approaches used for standard sequence-analysis tasks and the key parameter choices that a user must make.

2.1. Nucleotide-Level Analysis Techniques

Because there are four nucleotide bases (A,T,C,G), the analysis of nucleotide sequences is generally much less complex than that of protein sequences.

2.1.1. DNA Sequence Alignment

The development of robust algorithms for the alignment of DNA sequences to one another might be considered the origin of bioinformatics. Although work on modeling metabolic networks predates sequence-alignment research, sequence-alignment algorithms have become integral components of molecular biology research. **Table 2** provides a listing of some key online resources for sequence alignment.

Sequence alignment refers to the matching of two sequences, letter-by-letter, with one another. This is commonly done by using a quantitative measure of "sequence-similarity" that is encoded in a "substitution matrix". This matrix indicates the scoring penalty to be assigned when two non-identical bases are aligned together, and the reward to be assigned when two identical bases are matched. An algorithm is then used to find "optimal scoring regions", which contain many alignments of identical bases and few alignments of dissimilar ones. Sequence alignment algorithms have been reviewed in a very accessible manner in several texts ***(14,15)***. Sequence alignments can either handle pairs of sequences, or many sequences simultaneously (**Fig. 3**). Here, we only discuss the simpler case of pair-wise-sequence alignments, but note that several reviews of multiple sequence-alignment algorithms exist ***(16–18)***.

Table 2
Selected tools for sequence alignment

URL	Notes
www.ncbi.nlm.nih.gov/BLAST	Gateway to NCBI's comprehensive BLAST resources. Suitable for all types of alignments.
genome.ucsc.edu/cgi-bin/hgBlat	UCSC's BLAT server is ideal for aligning ESTs/mRNAs onto genomic assemblies.
www.ebi.ac.uk/Tools/similarity.html	Gateway to ENSEMBL's comprehensive sequence-alignment resources. Similar to those offered by NCBI.
www.ebi.ac.uk/embl_services/index.html	Gateway to EMBL's sequence-analysis options, which include many more tools than NCBI or ENSEMBL.

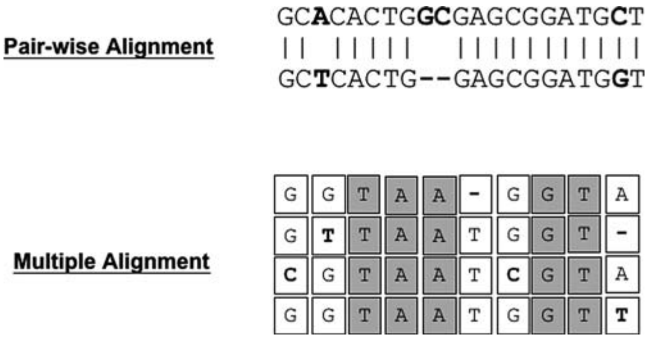


Fig. 3. Sequence-alignments can be divided into two groups based on the number of sequences in the alignment. Pair-wise alignments involve only two sequences, and are much less computationally demanding than multiple sequence-alignments. In general, conserved bases are highlighted by *shading* in multiple-alignments.

Pair-wise sequence alignments can be divided into two basic classes: global alignments, which attempt to align complete sequences, and local alignments, which look for short, well-matched regions (**Fig. 4**). By far the most common sequence-alignment program is BLAST (basic local alignment search tool), which, as its name indicates, performs local alignments. The popularity of BLAST is largely a result of two factors. First, it is a very rapid algorithm, capable of finding relevant matches in minutes. Second, because

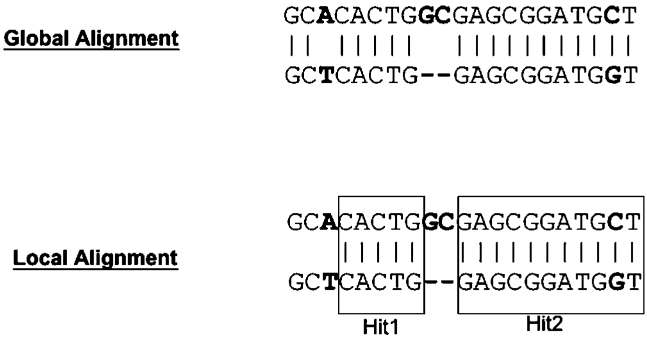


Fig. 4. Sequence alignments can be divided into two groups based on the extent to which the alignment covers the sequence. If an alignment covers the entire extent of a sequence, the alignment is called a "global" one, whereas if it is broken up into local regions of higher similarity, often called "hits," this would be a local alignment.

NCBI researchers developed BLAST, it has been extensively integrated into the NCBI web site, and has been given a very user-friendly interface. Indeed, BLAST is actively developed and has undergone several improvements since its original publication ***(19)***, such as adding the ability to seamlessly handle gaps within a sequence ***(20)***, more accurate statistical estimation ***(21,22)***, and addition of versions of BLAST specialized for specific tasks ***(23)***.

While many other sequence-alignment algorithms outside of BLAST ***(24, 25)*** exist, the major parameter-selection choices for all pair-wise alignment algorithms are similar. First, and most critically, the user must select a database to query for this alignment. The default database options are typically very liberal, and often include sequences from multiple species. For many gene-search queries, for example, a species-specific EST database is more appropriate than a database containing genomic and non-genomic sequences from multiple organisms. Selecting an appropriate database reduces the number of spurious matches, improves statistical significance, and thus reduces execution time. Second, and equally important, a user often needs to select a "sensitivity" parameter – this is called "word-size" for BLAST algorithms, and "ktup" for FASTA algorithms. In general, lower values of these parameters ensure that fewer hits are missed by the algorithms at the cost of increased execution time. Detailed guidelines for choosing BLAST parameters are also available ***(26,27)***. *See* **Chapter 13** for more discussion on BLAST.

2.1.2. DNA Promoter Analysis

While experimental techniques to determine the genomic binding-sites of transcription-factors are now becoming available ***(28–30)***, computational approaches remain a common way of developing hypotheses about the regulation of a given gene. In particular, lists of genes found to be co-expressed in microarray experiments are commonly subjected to word-search or word-enrichment analyses to identify novel or known transcription factors that might be contributing to this regulation ***(31,32)***. For a listing of some key online tools for promoter analysis *see* **Table 3**.

These tools generally follow one of three major approaches: (1) Phylogenetic footprinting, (2) word-frequency analysis, and (3) library-based motif-searching. Phylogenetic footprinting exploits evolutionarily relationships to identify well-conserved regions of DNA – such regions have been demonstrated to be enriched for transcription-factor binding-sites ***(33,34)***. By contrast, motif-searching analyses look for the over-representation of short DNA strings in a series of genes, such as those identified by microarray analysis ***(35–39)***. While word-enrichment searches can often identify novel transcription-factor

Table 3
Selected tools for promoter analysis

URL	Notes
bioportal.bic.nus.edu.sg/tres/	The TRES* database promoter search site offers multiple analysis tools through a single interface.
www.cisreg.ca	Listing of software tools from the Wasserman laboratory—one of the leaders in prediction and analysis of TFBS** data. Includes detailed presentation slides.
jaspar.cgb.ki.se	The JASPAR database is a free listing of known TFBS sequences.

* Transcription regulatory element search.
** Transcription factor binding site.

binding-sites, library-based searches exploit databases of known transcription-factor specificities *(40)*. As such, library-searches are generally computationally rapid, although sometimes less useful for knowledge-discovery processes.

Parameter choice for promoter-analysis is generally straightforward as selection of a stringency score or a *P*-value threshold is typically sufficient. This single parameter will often be the primary control for the sensitivity/specificity trade-off for a given analysis.

2.1.3. DNA Primer Design

Polymerase chain reaction (PCR) analysis is another key component of modern molecular biology. Because a large fraction of PCR specificity and sensitivity is dependent on the primer sequences, an array of different computational approaches have been developed to help improve selection of optimal primers. One of the most common programs, Primer3, provides detailed biophysical modeling of primer characteristics ***(41)***. In addition, databases of primer sequences have been developed ***(42)*** to allow users to leverage already-tested PCR data more efficiently. *See* **Table 4** for a listing of some key online resources for primer-design.

More recently a number of groups have developed BLAST-based techniques for validating primer-accuracy, particularly in the context of validating mRNA expression microarray data. These programs, such as VPCR ***(43)*** and PUNS ***(44)*** allow the user to determine rapidly if their primer pair will amplify multiple products at various stringency levels. This ability to simulate the results of a PCR reaction can be very powerful. For example, when reverse transcriptase

Table 4
Selected tools for primer design

URL	Notes
frodo.wi.mit.edu	Primer3 is one of the most sophisticated, flexible primer-design tools available.
okeylabimac.med.utoronto.ca/PUNS/	PUNS provides a flexible, multi-species method for assessing primer-specificity.

PCR is used for assessing mRNA levels it is important to rule out the possibility of genomic contamination. This is typically done by selecting primers that span intron-exon boundaries. PCR simulation tools allow this to be done in a rapid fashion, allowing genomic contamination to be ruled out for specific pairs of primers.

2.1.4. RNA Analysis

Although DNA can be considered as primarily a linear chain, RNA can fold into complex three-dimensional shapes—shapes which often have functional roles ***(45)***. As a result RNA analysis can be considered as a hybrid between three-dimensional protein analysis and linear DNA analysis. For example, an extensive study of the 3’ and 5’ untranslated regions (UTRs) of human mRNAs revealed a variety of novel sequences that appear to be associated with RNA binding proteins, and may be functionally relevant ***(46)***. For a listing of some key tools for RNA analysis, *see* **Table 5**.

Table 5
Selected tools for RNA analysis

URL	Notes
www.bioinfo.rpi.edu/˜zukerm/rna/	Home-page for the Zuker lab – a leading group in the study of RNA folding, and developers of the popular *mfold* program.
www.tbi.univie.ac.at/˜ivo/RNA/	The Vienna package of RNA structure prediction and comparison tools is a comprehensive, but advanced, set of tools.

2.1.5. RNA Folding

RNA folding can be assessed by a variety of different techniques, but in general is much simpler to estimate than is protein folding. The RNA folding problem can be simplified to that of determining which base-pairs are "paired," and which are "unpaired" ***(47)***. Groups of paired bases form "beads" of structure surrounded by flexible, unordered single-stranded RNA. These beads fall into four general secondary-structure classes: helices, loops, bulges, and functions ***(48)***.

While prediction can be computationally demanding, the tractability of this problem is apparent in the diversity of approaches, each of which can achieve relatively high (e.g., ~70%) accuracies. The four major classes of algorithms used to predict the three-dimensional shape of RNAs are called the "naïve," "empirical," "homology-based," and "thermodynamic" algorithms. These algorithms are mainly distinguished by their simplifying assumptions. For example, naïve approaches exploit statistics from known structures to predict new structures. By contrast, thermodynamic approaches assume that the process of folding involves a descent down a free energy well, to a maximally stable conformation. A variety of web-servers exist for exploiting these algorithms, each with their own specific parameters. Naïve algorithms generally take a probability threshold; while thermodynamic algorithms require determination of how "quickly" the structure might be expected to reach its most stable state i.e. cooling temperature. In general these methods are highly sensitive to parameter choice, and careful review of each individual program is necessary to obtain accurate results ***(49)***.

2.2. Protein-Level Analysis Techniques

Analysis of proteins is generally much more complex than the analysis of nucleotide sequences. This is not only a result of the folding of proteins into complex three-dimensional shapes, but also from the greater complexity of the 20-letter amino-acid alphabet relative to the simpler, more chemically homogeneous nucleotide alphabets. While proteins can be analyzed with motif-searches, much like those used in promoter-analysis ***(50)***, the focus here will be on the analysis of protein structure and protein-protein interactions.

2.2.1. Protein Structure

The prediction of protein structures *de novo*, from the primary structure amino-acid sequence alone, remains an active area of investigation. One of the major problems in predicting protein structure is the immense computational complexity involved in estimating the structure of even a short peptide. Each individual side-chain to side-chain or side-chain to backbone interaction must be modeled, and can have effects that propagate to spatially distant regions of the protein. Indeed, it is just this physical flexibility that gives proteins much of their biological versatility, but it makes computational analysis intractable with brute-force algorithms. A listing of some key tools for the analysis and visualization of protein structures is found in **Table 6**.

For protein structure analysis, three sophisticated approaches are employed. The first uses so-called "comparative" algorithms to generate complete structural models based on sequence-similarity to a protein with an experimentally known structure. Unfortunately, a sequence-similarity of at least 30%, and possibly as high as 50%, is thought to be needed for accurate application of this method, making it inappropriate for most proteins ***(51)***.

The second approach involves de novo analysis to avoid this problem entirely, and predict three-dimensional structure directly from primary sequence. These methods are based on the assumption that proteins will generally fold into maximally stable states, and are thus related to "thermodynamic" RNA folding algorithms. The ROSETTA web-server is a major resource for this type of analysis ***(52)***.

The last approach for predicting protein structures involves "threading" algorithms. Like the comparative algorithms, threading methods exploit databases of solved structures. In this case, the sequence of the protein under investigation is "threaded" through known crystal structures. The energetic stability of the resulting alignments is calculated, and used to determine the

Table 6
Selected tools for protein structure analysis

URL	Notes
bmerc-www.bu.edu/psa/	The Protein Sequence Analysis server predicts secondary structure based on primary sequence.
www.ebi.ac.uk/Tools/structural.html	EBI provides an extensive of structural-analysis tools
www.rcsb.org/pdb/	The Protein Data Bank is the standard repository for solved three-dimensional structures.
ca.expasy.org/spdbv/	The Swiss PDB Viewer is an outstanding tool for working with three-dimensional structures.

accuracy of the threading ***(51)***. With increasing numbers of structures being solved, threading methods are gaining in popularity.

2.2.2. Protein–Protein Interactions

An emerging area of research is the prediction, characterization, and integration of protein-protein interactions. These interactions have been determined by a variety of high-throughput experimental techniques, such as mass spectroscopy ***(53)*** and yeast-two-hybrid approaches ***(54)***. A variety of computational approaches have been developed to identify novel protein-protein interactions. These range from phylogenetically-based techniques like gene-fusion analysis ***(55)*** and cross-species extrapolations ***(56)***, to techniques exploiting other sources of data like co-expression analysis ***(57)***, co-localization ***(58)***, or co-essentiality ***(59)***. *See* **Table 7** for a listing of some key online resources for protein-protein interaction analysis.

With such a diverse array of methods for predicting protein-protein interactions computationally or discovering them experimentally in high-throughput approaches, the organization and analysis of protein-protein interaction networks has become a major field of research. A series of databases have been developed to store known or predicted protein interactions ***(60)***, some of which include annotation to computational sources of support for these interactions ***(61)***. Similarly, several groups have looked at motifs within these networks ***(62)*** in an effort to determine whether there are broad classes of functional regulation that can be identified and investigated using a similar battery of experimental techniques.

While this type of work on validating and storing of protein-protein interactions is ongoing, effort is now underway to exploit these interaction networks and to use them in the context of broader biological investigations. For example, an extensive integration of ChIP-chip (transcription-factor binding), mRNA

Table 7
Selected tools for protein-protein interactions (PPIs)

URL	Notes
ophid.utoronto.ca	OPHID is a comprehensive database of both known and predicted human protein-protein interactions. It combines these interactions with sources of computational evidence.
www.hprd.org	The Human Protein Reference Database is an extensive listing of literature-curated human PPIs.

expression, and protein-interaction data was recently published ***(63)***. Similarly, genes whose expression is perturbed in lung cancer have been shown to occupy central, well-connected positions in protein-protein interaction networks ***(64)***.

References

1. Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington D.C.
2. Baxeveanis, A. D. O., and Ouellette, B. F. F. (eds) (2005) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 3rd ed. Wiley-InterScience, Hoboken, NJ.
3. Bateman, A. (2006) Editorial, *Nucleic Acids Res*. **34**, Database Issue 1.
4. Geer, R. C., and Sayers, E. W. (2003) Entrez: making use of its power. *Brief Bioinform*. **4**, 179–184.
5. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res*. **33**, D54–D58.
6. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2003) NCBI reference Sequence project: update and current status. *Nucleic Acids Res*. **31**, 34–37.
7. Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., et al. (2004) Ensembl 2004. *Nucleic Acids Res*. **32**, D468–D470.
8. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002) The human genome browser at UCSC. *Genome Res*. **12**, 996–1006.
9. Openhelix, http://www.openhelix.com/ucscmaterials.shtml.
10. Safran, M., Chalifa-Caspi, V., Shmueli. O., Olender. T., Lapidot, M., Rosen, N., et al. (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*. **31**, 142–146.
11. Rosen, N., Chalifa-Caspi, V., Shmueli, O., Adato, A., Lapidot, M., Stampnitzky, J., et al. (2003) GeneLoc: exon-based integration of human genome maps. *Bioinformatics* **19**(S1), 222–224.
12. Lenhard, B., Hayes, W. S., and Wasserman, W. W. (2001) GeneLynx: a gene-centric portal to the human genome. *Genome Res*. **11**, 2151–2157.
13. Gilbert, D. G., (2002) euGenes: a eukaryote genome information system. *Nucleic Acids Res*. **30**, 145–148.
14. Gribskov, M. R., and Devereux, J. (1991) Sequence analysis primer. UWBC biotechnical resource series, Stockton Press; Macmillan Publishers, New York, xv, 279, .
15. Durbin, R. (1998) Biological sequence analysis: probabalistic models of proteins and nucleic acids, Cambridge University Press, Cambridge, p. 356.
16. Thompson, J. D., Plewniak, F., and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*. **27**, 2682–2690.
17. Phillips, A., Janies, D., and Wheeler, W. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol*. **16**, 317–330.

18. Baldauf, S. L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19**, 345–351.
19. 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.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
21. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005.
22. States, D. J., Gish, W., and Altschul, S. F. (1991) Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods: A Companion to Methods in Enzymology* **3**, 66–70.
23. Tatusova, T. A., and Madden, T. L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250.
24. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., et al. (2000) PipMaker—web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586.
25. Kent, W. J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
26. Pertsemlidis, A., and Fondon III, J. W.. (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol.* **2**, Reviews 1–10
27. Boutros, P. C. (2005) An Introduction to Effective BLASTing. *Hypothesis* **3**, 26–33.
28. Liu, X., Noll, D. M., Lieb, J. D., and Clarke, N. D. (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* **15**, 421–427.
29. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* **298**, 799–804.
30. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., et al. (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309.
31. Frith, M. C., Fu, Y., Yu, L., Chen, J. F., Hansen, U., and Weng, Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* **32**, 1372–1381.
32. Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23**, 137–144.
33. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E., (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**, 225–228.

34. Boutros, P. C., Moffat, I. D., Franc, M. A., Tijet, N., Tuomisto, J., Pohjanvirta, R., et al. (2004) Dioxin-responsive AHRE-II gene battery: identification by phylogenetic footprinting. *Biochem. Biophys. Res. Commun.* **321**, 707–715.
35. Zhu, Z., Pilpel, Y., and Church, G. M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.* **318**, 71–81.
36. Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics* **3**, 34.
37. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176.
38. Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805.
39. Mwangi, M. M., and Siggia, E. D. (2003) Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics* **4**, 18.
40. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94.
41. Rozen, S., and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386.
42. Pattyn, F., Speleman, F., De Paepe, A., and Vandesompele, J. (2003) RTPrimerDB: the real-time PCR primer and probe database. *Nucleic Acids Res.* **31**, 122–123.
43. Lexa, M., Horak, J., and Brzobohaty, B. (2001) Virtual PCR. *Bioinformatics* **17**, 192–193.
44. Boutros, P. C., and Okey, A. B. (2004) PUNS: transcriptomic- and genomic-in silico PCR for enhanced primer design. *Bioinformatics* **20**, 2399–2400.
45. Moore, M. J. (2005) From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309**, 1514–1518.
46. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345.
47. Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A., and Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29**, 4724–4235.
48. Tinoco, I., Jr., and Bustamante, C. (1999) How RNA folds. *J. Mol. Biol.* **293**, 271–281.
49. Major, F., and Griffey, R. (2001) Computational methods for RNA structure determination. *Curr. Opin. Struct. Biol.* **11**, 282–286.
50. Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., et al. (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**, 383–387.

51. Baker, D., and Sali, A. (2001) Protein structure prediction and structural genomics. *Science* **294**, 93–96.
52. Meiler, J., and Baker, D. (2003) Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci. U S A* **100**, 12105–1210.
53. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207.
54. Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324.
55. Tsoka, S., and Ouzounis, C. A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat. Genet.* **26**, 141–142.
56. Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., et al. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U S A* **102**, 1974–1979.
57. Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A., et al. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell.* **9**, 1133–1143.
58. Huh W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., et al. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691.
59. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., et al. (2003) Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453.
60. Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., et al. (2006) Human protein reference database—2006 update. *Nucleic Acids Res.* **34**, D411–D414.
61. Brown, K. R., and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics* **1**, 2076–2082.
62. Przulj, N., Wigle, D. A., and Jurisica, I. (2004) Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340–348.
63. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312.
64. Wachi, S., Yoneda., K., and Wu, R. (2005) Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205–4208.

18

Developing Decision Support Systems in Clinical Bioinformatics

Vitali Sintchenko and Enrico Coiera

Summary

There is a growing demand for tools to support clinicians utilize genomic results generated by molecular diagnostic and cytogenetic methods in support of their decision-making. This chapter reviews existing experience and methods for the design, implementation and evaluation of clinical bioinformatics electronic decision support systems (EDSS). It provides a roadmap for identifying decision tasks for automation and selecting optimal tools for building task-specific systems. Key success factors for EDSS implementation and evaluation are also outlined.

Key Words: decision support systems, decision-making, clinical bioinformatics; genomics; artificial intelligence; risk assessment.

Abbreviations: EDSS – electronic decision support systems; LOINC – Logical Observation Identifier Names and Codes; ROC – receiver-operating characteristic curve; SNOMED® — Systematized Nomenclature of Medicine; UMLS – United Medical Language System

1. Introduction

Computerized decision support systems have become an essential part of the vision of evidence-based decision-making, aimed at enhancing the quality and effectiveness of clinical decisions ***(1,2)***. The clinical decision process is challenged by the amount of clinical data now available, and the expanding knowledge base generated by new technologies and clinical trials. For example, there are estimates that in just a few years, primary care practitioners will have to know how to employ as many as 100,000 new genetic screening tests ***(3)***.

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

Decision aids can significantly reduce human error and have been advocated as a mechanism for the translation of genomics, proteomics, transcriptomics, and metabolomics into new clinical decision models, leading to more personalized medical approaches *(3)*. Decision aids with a clinical bioinformatics focus have been recently developed including patient-specific risk assessment tools with potential for early warning, risk prediction and assessment, and treatment follow-up ***(5–7)***. They target the range of monogenic inherited disorders, somatic mutations and gene expression profiling as well as complex multifactorial disorders *(8)*. For example, personalized risk calculators for breast cancer (*see* **Note 1**) and preoperative complications based on genomic data have been developed ***(5,9,10)*** (*see* **Note 2**). They also notify clinicians when their patients might be eligible for a pertinent clinical trial based on either their genotypic or phenotypic patient characteristics *(3)*.

We define electronic decision support systems (EDSS) as tools that provide access to knowledge stored electronically, and that aid clinicians in making decisions. They encompass a variety of systems and interventions such as computerized alerts and reminders, expert systems, electronic clinical guidelines, practice protocols, pathology order sets, and clinical workflow tools. Software designed to support biomedical research tasks such as sequence similarity and alignment assessment, gene or protein discovery and prediction, and genetic classification and automated sub-typing algorithms have been reviewed elsewhere ***(11,12)*** and will not be considered here (*see* also **Chapter 17**).

EDSS in clinical bioinformatics do differ from traditional decision aids in some ways, usually because they focus either on new clinical tasks or

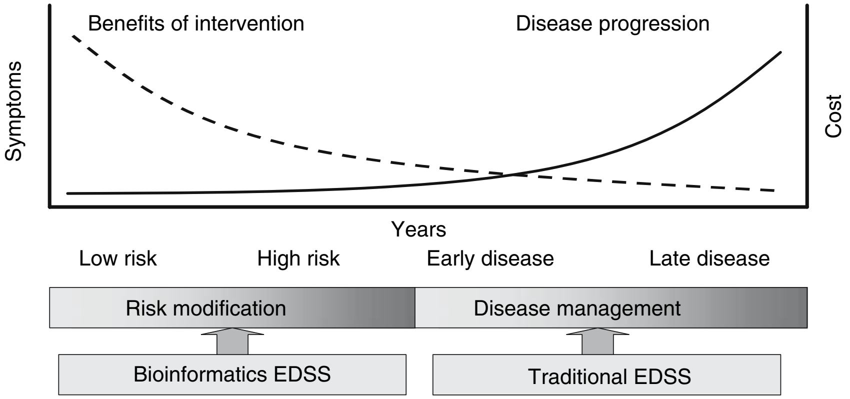


Fig. 1. Risk assessment decision support.

new types of information. Specifically, they may address decisions related to early detection and prognosis of diseases at the pre-symptomatic stage (**Fig. 1**) and utilize risk calculations based on genomic, proteomic or transcriptomic data. Such decisions are often bound by significant uncertainty, they are time-consuming, and clinicians are unlikely to be familiar with these tasks.

Clinical bioinformatics EDSS, in contrast to conventional EDSS, can enhance our capacity for early detection and treatment allowing time for preventative interventions. For example, the assessment of alternatives is assisted by calculation of patient-specific risks of diseases with a large genetic component or outcomes associated with the carriage of genes with high penetrance and processing complex molecular typing patterns and issuing clonal alerts when matching genotypes are detected.

Examples of task-specific clinical decision support systems in use are listed in **Table 1**. Cancer prognostics has been one of the first test cases for bioinformatics EDSS, given the fact that cancer is caused by genomic instability, and

Table 1
Task-specific decision support systems in clinical bioinformatics

Task	Information Support	Examples of Systems
Provision of information relevant to the decision to assess alternatives	Evidence-based information about options and chances of different outcomes occurring with these options Education and decision counseling	Risk Assessment in Genetics (RAG) *(5)* Breast cancer management decision aid *(13)*
Help with the structuring of a decision and preference clarification	Information about diagnostic biomarkers and biomarkers of disease progression Information about personal risk levels	AdjuvantOnline www.adjuvantonline.com
Processing of the information	Calculation and/visualization of patient- or population-specific risks Choice of the 'best' option e.g., the most cost-effective one	HIV genotypic resistance test interpretation systems *(7)* Biosurveillance alerts (identification of molecular clusters) *(14)*

microarrays potentially allow assessment of patients' entire expressed genomes. Analysis of breast cancer patients' expression patterns can already be highly correlated with recurrence risks ***(15)***. Family breast cancer risk assessment tools to estimate patient susceptibility, survivability and recurrence have been employed to identify individuals at high risk of cancer who may benefit from targeted screening or prophylaxis, e.g., tamoxifen chemoprevention for women aged 35 or older with a 1.67% or higher 5-year breast cancer risk cutoff calculated on the Gail model ***(9)***. Evidence suggests that EDSS can successfully support tasks related to clinical decisions associated with genomic medicine by providing relevant information at the point of decision-making ***(13,16)***.

2. System Design

2.1. Choice of Tasks Suitable for Automation

The design of a clinical EDSS begins with the characterization of a decision task, and includes identifying the available data, the available knowledge to guide the decision process, the setting in which the decision is made, the decision maker's specific needs and resources, the task's informational structure and the specific information needs of defined subtasks such as data input. Failure to adequately characterize the task to be supported is a common cause of poor system performance once deployed in a working setting, independent of the quality of the software system itself. Indeed more than half the errors which occur during systems development may be due to requirements errors (where the requirements specification does not match actual user requirements) ***(3)***.

Practitioners with different training and clinical roles may prefer quite different tools to optimize their decisions. For example, a primary care practitioner (also called general practitioner, or family physician—*see* **Chapter 19**) dealing with a patient anxious about her breast cancer family risks will probably need a very different tool compared to that required by a specialist surgeon advising the same patient about her treatment options. The uptake of EDSS is also influenced by the attitudes of decision-makers. There is significant variability in personal beliefs and preferences for evidence seeking and decision support between different clinical professional groups and individual clinicians.

Decision support is especially relevant for tasks that are cognitively demanding, routine and high volume, or are error-prone or infrequent but have important outcomes. Increasing complexity of a decision process is likely to be associated with an increased risk of human error, either because the decision task exceeds inbuilt human cognitive limits such as short term memory, or because of work-arounds or heuristics which attempt to simplify the task but result in poorer decision outcomes. A corollary is that EDSS are unlikely to be adopted in situations in which they impose additional workload but deliver minimal additional benefit, e.g., for routine clinical decision processes which are well understood, are of minimal complexity and impose little cognitive load. Traditionally areas of high adoption for EDSS include clinical laboratories, where decision volumes are high, or in medication prescription support, where the complexity and risk of drug-drug interactions is such that unassisted prescribing becomes an unacceptable and unsafe clinical practice.

If decision support does not reduce a complex task into a simple one, without loss of decision quality, then the performance of the task is unlikely to benefit from automation. Complexity of a task is thus a central feature in determining EDSS success ***(16,17)***. From the perspective of information theory, task complexity measures the amount and structure of the information that needs to be processed. Complex tasks may have a large number of subtasks, inputs and products with elements that are probabilistic in their behavior and may evolve over time. The process of decision-making and flow of associated data are often represented in functional specifications as Data-Flow Diagrams (*see* **Fig. 2** for an example). Decision complexity can be assessed by one or a combination of approaches, e.g., minimum length of the message ***(18)***, evaluation of cognitive effort ***(19)***, and Clinical Algorithm Score ***(20)***. To decide whether automation will benefit a task, the following stages have been suggested as a good filtering process:

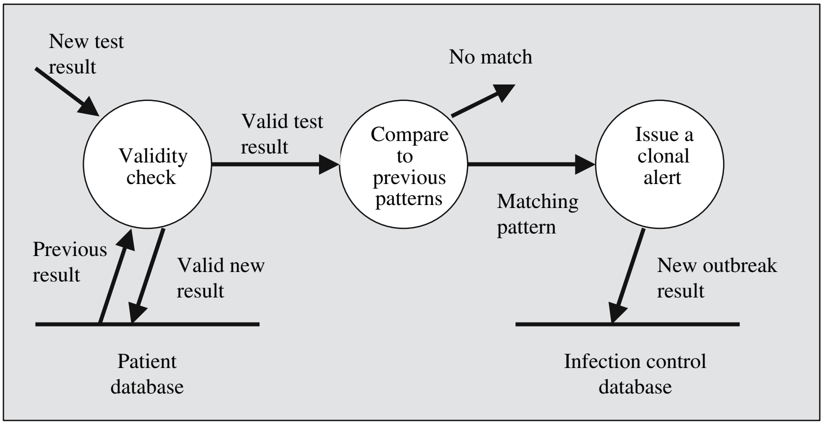


Fig. 2. A data-flow diagram that graphically represents the process and data flows within a biosurveillance system. Bubbles depict processes, vectors depict data flows, and straight lines depict databases.

1. Select the domain and decision tasks
2. Evaluate the complexity of knowledge required for the clinical tasks selected
3. Select the (potentially) most cognitively demanding tasks based upon the comparison of their complexity
4. Assess unaided and EDSS-aided cognitive effort for the selected tasks, to determine if complexity reduction is possible with the use of an EDSS
5. Select computational tools to achieve reduction of task complexity for the user

Sintchenko and Coiera ***(17)*** provide more details on the specific methods for task complexity assessment.

2.2. Building the EDSS

2.2.1. Components of an EDSS

A decision support system at its most abstract encodes one or more *decision procedures* within a *knowledgebase*, and based upon data presented to it by a *database*, draws inferences based upon a predefined set of *decision rules*. The knowledgebase is essentially a store of decision procedures, which is used to generate the EDSS recommendations (**Fig. 3**). For example, a set of if-then rules might be used to encode which diagnosis is most likely based upon the presence or absence of patient data.

The decision rules are the methods used to match the knowledgebase to the database, and are typically either the laws of probability, e.g., when the EDSS is required to make suggestions based upon likelihoods, or the rules of logic as might occur when knowledge is encoded as a set of if-then rules. Other well-known decision methods include neural nets and decision trees (*see* ***(8)*** for more details). The level of accuracy needed for a prediction rule to be clinically

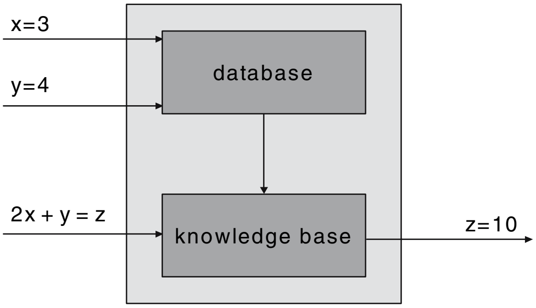


Fig. 3. A decision support system encodes one or more decision procedures within a *knowledge base*, and based upon data presented to it by a *database*, draws inferences based upon a predefined set of *decision rules* ***(8)***.

useful is more stringent that necessary for determining that gene expression profiles significantly differ between two groups. For example, a predicted 70% recurrence probability should be treated quite differently by clinicians if the associated uncertainty is 30%, than if it were 2%.

The challenge for most EDSS is the process of building the knowledge base. Traditionally there have been two separate processes available. In well-understood domains, where human experts are available to articulate the decision procedures, the knowledge base can be hand crafted using one of several different knowledge elicitation procedures. Perhaps the most widely used and robust approach to hand crafted knowledge based development is the ripple down rule (RDR) approach in which experts provide rules to classify data sets such as laboratory results, and refine the knowledge base only when the initial rule set fails ***(21)***. In domains where knowledge is less explicitly modeled, then automated methods for knowledge base construction are favored.

2.2.2. Automated Knowledge Base Development

Machine learning or data mining methods are of particular interest in clinical bioinformatics, where explicit knowledge is scarce or rapidly evolving, but where there are large data sets which can be processed to discover likely relationships between clinical conditions and biological markers. A wealth of literature describes computational techniques to discover and explore quantitative associations between classes or clusters and to generate semantic descriptions of clinical categories, such as types of disease or prognostic conditions ***(6,22–24)***. Most such methods include a training phase run on samples whose classes are already known, and a testing phase, in which algorithm generalizes from the training data to predict classification of new samples (**Fig. 4**). Because of this directed training phase, prediction methods are referred to as "supervised" classification methods.

For genomic or proteomic data, prediction generally refers to the classification of patients' samples by characteristics such as disease subtype or response to treatment ***(24,25)***. Choosing a prediction method requires selecting from a vast range of techniques. Conventional linear discriminant methods have been extended to include weighted voting ***(26)***, shrunken centroids ***(27)***, and compound covariates ***(28)***. Powerful machine learning approaches are also k-nearest neighbor prediction and neural networks ***(24)***. Two other classes of algorithms are of growing interest for multidimensional learning problems: support vector machines and decision tree classifiers ***(29,30)***. The number of classes in the prediction problem and small sample size may impose additional constraints on the choice of algorithms. Whereas decision trees, neural networks

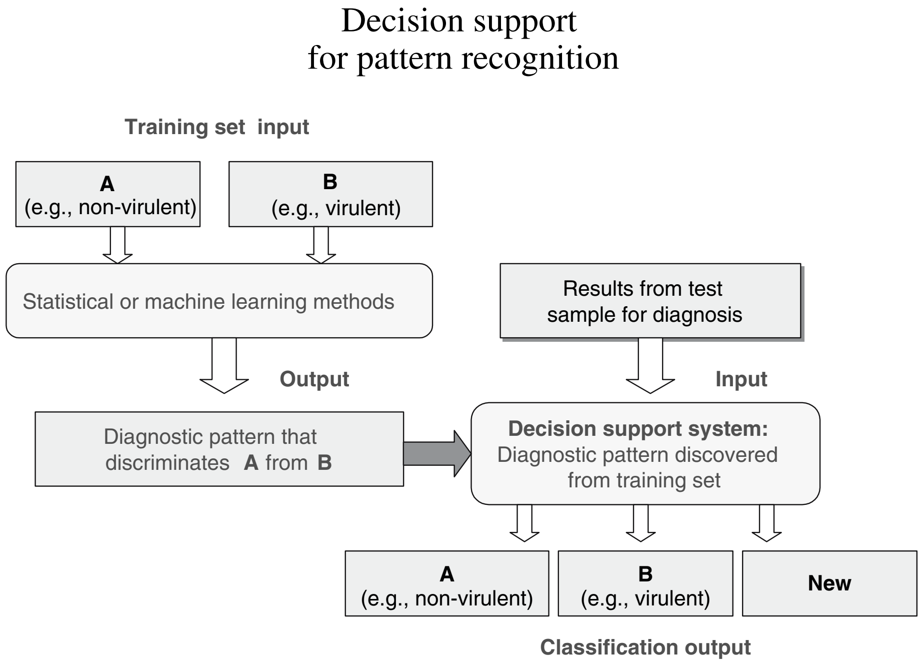


Fig. 4. Classification of data using machine learning approaches.

and k-nearest neighbors can, in principle, separate any number of output classes, support vector machines and linear methods are inherently binary.

There is no universal pattern recognition or classification model to predict molecular profiles across different data sets and medical domains. Many classification and knowledge discovery problems may require the combination of multiple techniques not only to improve the accuracy and efficiency of the analysis task (**Table 2**), but also to support evaluation procedures ***(24)***. There are several tools that integrate open and scalable research platforms, e.g., WEKA—Waikato Environment for Knowledge Analysis ***(40)***.

Table 2 outlines the main stages in preparation of a training data set for use by a machine learning algorithm. Data invariably require some preprocessing to "clean" it of noise, ensure that classification labels are applied consistently to all examples within the data set, and often will require some attempt to identify the features within the data set most likely to be associated with the biological phenomenon of interest. Whilst some algorithms will look for the most useful features, others will benefit from the use of human domain expertise in selecting a useful subset of the full feature set for learning. Simultaneous consideration

Table 2
Machine learning scheme

Step 1. Preprocessing	
Objectives	Removal of irrelevant or redundant data, noise reduction and normalization of the data from different samples
Methods	1. Heuristic noise reduction, e.g., smoothing filters, the wavelet transform 2. Model-based noise reduction
Comments	1. Heuristic noise reduction - Adding irrelevant attributes reduces the performance of decision trees and rules, linear regression and clustering methods (31,32,40). 2. Model-based noise reduction —essential if the task involves numerical attributes but the chosen method can only handle categorical ones ***(33)***
Step 2. Feature Extraction	
Objectives	Extraction of attributes corresponding to distinct pathological states
Methods	1. Attributes from original space ***(31)*** 2. Projecting signals into a lower-dimensional space using linear transformation, e.g., principal component analysis
Comments	1. Projecting signals—Principal component analysis (PCA) identifies the orthogonal directions in which data vary maximally. Very sensitive to the choice of vectors thus criteria for selecting vectors should be determined prior to feature extraction (34,35,40).
Step 3. Feature Selection	
Objectives	Reduction of dimensionality of the data and increase the likelihood of successful classification
Methods	1. Filter method 2. Wrapper method 3. Embedded methods
Comments	1. Filter method —Independent assessment based on general characteristics of data. Determine the subset for classification by ranking individual features based on selection criteria, e.g., t statistics ***(36)***. 2. Wrapper method—The learning algorithm is wrapped into the selection method. Determine the subset for classification by evaluating the relevancy based on metrics of a classifier trained using the subset of features. ROC analysis can be used to measure the relevancy of individual attributes (31,37).

(*Continued)*

**Table 2
(Continued)**

	3. Embedded method—Implicitly perform feature selection as a part of the classifier training process, e.g., decision trees (35,36).
Step 4 Classifier Training	
Objectives	Distinguish classes based on selected features
Methods	1. Unsupervised machine learning or clustering 2. Supervised machine learning
Comments	1. Unsupervised machine learning or clustering—Natural groupings are identified based without predefined “correct” class membership examples, e.g., hierachical clustering algorithms, self-organizing maps ***(36)***. 2. Supervised machine learning—Classifier is developed using a subset of data with predetermined classes, e.g., artificial neural networks, k nearest neighbor, linear discriminant analysis, support vector machine, Naïve Bayes, rule induction etc (32,38–40).
Step 5 Classifier Evaluation	
Objectives	Assess the performance of a classifier
Comments	Ideally, separate data sets should be used for stages 4 and 5. In practice, however, data partitioning of a single data set, such as 10-fold cross-validation or bootstrap sampling are employed for small size datasets. Over-sampling the minority class and under-sampling the majority class have been common methods to resolve biased classification due to imbalanced data (31,35).

of features, e.g., a composite medical index or panel of markers, may provide more information than individual indicators because the predictability of an outcome is based not on presence or absence of several biomarkers or their linear summation, but on a complex, non-linear relationship between them.

2.2.3. Standards for Data Integration

Information which is relevant to genomic profiling exists in a variety of sources and formats. For EDSS which are “home grown” using local data, and which will only have local institutional use, there may be no compelling reason to adopt a standardized approach to representing data. However, there is an increasing focus on linking disparate databases, disease

Table 3
Standards for data representation and storing

Standards	Examples	URL / Reference
Knowledge engineering standards	CommonKADS	www.commonkads.uva.nl
	OIL	www.ontoknowledge.org
	OML	www.ontologos.org/oml
	Knowledge Query and Manipulation Language	www.cs.umbc.edu/kqml
Software engineering standards	Case Data Interchange Format	(44,45)
	Information Resource Dictionary System	(44)
	Open Information Model	Microsoft
	Unified Modeling Language	(41)
WWW standards	XML	www.w3.org/XML
	Document Content Description	www.w3.org/tr/note-dcd
	Resource Description Framework	
	Web Ontology Language	www.w3.org/2001/sw/WebOnt
Bioinformatics standards	MAGE-ML Microarray and Gene Expression Markup Language	www.geml.org
	Clinical Bioinformatics Ontology	www.clinbioinformatics.org
	BioPathways Consortium	www.biopathways.org
	Gene Ontology Consortium	www.geneontology.org
Medical terminology	SNOMED-CT	www.snomed.org
	UMLS	41

registries, and clinical repositories, and for this to occur the task is substantially simplified if all data are represented in as uniform and standard a way as possible (**Table 3**). For example, databases of microbial genotyping results and clinical observations relating phenotype to genotype form an important part of the genetic variation data landscape. A compilation of microbial reference sequences (RefSeq) specifying gene name and DNA sequences can be found at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi (bacterial RefSeq), http://www.ncbi.nlm.nih.gov/genomes/FUNGI/funtab.html (fungal RefSeq) and http://www.ncbi.nlm.nih.gov/genomes/static/vis.html (viral RefSeq).

The ability to capture and share profiling data depends on shared use of a vocabulary (the words), syntax (the "sentence" structure), and messaging protocols. The most developed health care vocabularies are the United Medical Language System (UMLS, National Library of Medicine), LOINC (Logical Observation Identifier Names and Codes; Regenstrief Institute) and SNOMED® (Systematized Nomenclature of Medicine; College of American Pathologists) ***(42)***.

LOINC is an exhaustive catalogue of laboratory tests distinguished by source, e.g., serum or tissue, method, e.g., microscopy, PCR, or immunoassay, and the format in which the result is represented (ordinal, nominal or quantitative). The LOINC number describes a test, but does not provide the result of a specific test ***(42)***. In contrast, SNOMED® is a concept-oriented electronic vocabulary pioneered by the College of American Pathologists. SNOMED-Clinical Terminology (SNOMED-CT) contains around 364,000 concepts, 984,000 terms and 1.45 million defined relationships between concepts ***(43)***. It distinguishes concepts for a condition, e.g., haemochromatosis, the causative mutation, e.g., *BRCA1*, and diagnostic test, e.g., PCR. The UMLS maps the many different source terminologies available, and is a kind of terminological rosetta stone. It models individual systems, identifying for example the information about a laboratory test term, the source terminologies it comes from, which terms it is related to in the hierarchies of those source terminologies, what its synonyms and lexical forms are, and which other terms it is related to in some source terminology ***(43)***. It does not, however, strive to provide definitional information (such as what the test measures are or what its specimen is). However, synergistically, these vocabularies can support the integration of the high-level terms used in decision rules, e.g., "Haemochromatosis," with the relatively low-level terms used in the clinical records, e.g., "Blood test."

2.2.4. Socio-Technical Aspects of EDSS Implementation

The effective introduction and integration of new technology into existing processes requires user participation in design and interdisciplinary collaboration for iterative development. Decision making in healthcare is often more related to agreement with social expectations and the caretakers' perceptions of the clinicians' role than to standard biomedical rules. Therefore, a systematic approach to EDSS implementation, addressing characteristics of users, tasks as well as organizational context is usually fruitful. Specifically, implementation should take into account the differing needs of users with the variety of experience, training and clinical roles. System context also needs consideration, focusing on the situated conditions of use with explicit organizational goals, missions, control structures and communication modes. It is important to keep in mind potential professional, technical and personal barriers to uptake of EDSS (**Table 4**).

The implementation of EDSS faces the same barriers as the near-term diffusion of genomic medicine. Enthusiasm for the promise of genetic medicine on the part of medical geneticists contrasts markedly with the lack of relevant knowledge on the part of decision makers *(2)*.

3. Choice of Appropriate Evaluation Methodology

3.1. Evaluation Methodologies

Evaluation is central to any successful EDSS deployment, and should be conducted throughout the system development, starting at the planning and requirements stage and into implementation and the post release stages. Taking an iterative view of information system development, we can conceptually think of all these steps occurring within two different development cycles (**Fig. 5**):

1. Formative development cycle: The form that a system takes is iteratively determined by assessing user needs, designing prototypes, and then getting user feedback on system performance.
2. Summative assessment cycle: Once a system is robust enough for an outcomes assessment, it is put on trial and the summation of system performance results are used to drive the design of the next version of the system.

3.2. Formative Evaluation

At the formative stage of EDSS evaluation, performance of the system is assessed including accuracy of predictions, quality of sources, currency of knowledge and safety of recommendations. Iterative prototyping exposes small samples of prospective users and/or designers to a succession of evaluation protocols using simple models, storyboards, and interactive prototypes. Prototype evaluation uses qualitative methods such as cognitive walkthroughs, questionnaires, structured and informal interviews, focus group analyses, heuristic inspections, and verbal probes. Such evaluation should also include knowledge content evaluation with assessment of accuracy, sensitivity and specificity of classification methods, and estimating the optimum number of clusters to train genomic classifiers and learning parameters, as well as the selection of data sets, relevant features and classification models.

Table 4
Professional, technical, and personal barriers to usage of EDSS

Barriers	Examples	References
Barriers related to characteristics of EDSS		
Rule validity	1. Opinion-based recommendations 2. Insufficient cross-validation 3. Unproven cost-effectiveness	(46,47)
System relevance	1. Limited applicability to clinical practice, e.g., difference in patient mix. 2. Uncertainty about the “shelf-life“of EDSS	(48,49)
System practicality	1. Ambiguous output 2. Disruption to routine practice 3. Low uptake and clinical impact 4. Increase in consultation times	(8,50)
Barriers related to characteristics of EDSS implementation		
IT support	1. Lack of integration into existing systems 2. Lack of IT infrastructure	(47)
Insufficient evaluation	1. Lack of pre-implementation evaluation 2. Lack of post-implementation evaluation	(51)
Medico-legal concerns	No system for EDSS accreditation	(50)
Barriers related to characteristics of EDSS users		
Knowledge	1. Lack of awareness that quality of clinical decisions may be poor 2. Over-estimation of self-reported performance	(8,15)
Skills and abilities	1. Lack of IT skills 2. Belief that he/she cannot perform the task of EDSS use	(47)
Attitudes and beliefs	1. Low outcome expectations 2. Doubts about EDSS credibility 3. Uncertainty about medico-legal implications of EDSS use	(52,53)
Barriers related to characteristics of the organization or decision environment		
Established practices	1. Over-reliance on passive methods 2. Inertia of larger organizations	(47)
Culture	1. Resistance to change 2. Little or no history of EDSS use	(52)
Resources	1. Limited resources	(8)

Knowledge of organizational performance	1. Poor quality of clinical audit 2. Difficulty in measuring of outcomes 3. Short-term outlook rather than appreciation of long-term nature of EDSS impact and sustaining change	(47,51)
Patient factors	1. Preference over choices in clinical management	(51)

3.3. Summative Evaluation

A randomized, controlled trial is the ideal design for clinical impact analysis. Alternatives to a randomized trial include a "before-after" impact analysis (measures outcomes before, during and after using the EDSS) and an "on-off" impact analysis or interrupted time-series (measures outcomes in alternating time periods when the EDSS is or is not available). However, these designs are weaker, subject to temporal and "wash-over" confounding.

Assessment of outcome measures for EDSS should be blinded to patients' risk stratification and the decisions recommended by the EDSS. Ideally, this means that one group of clinicians use the DSS to make clinical decisions and a different group, unaware of the EDSS recommendations, assesses patients' clinical outcomes and impact measures. The potential for bias is significant when outcome events have subjective components.

Although a multi-institutional randomized study is the preferred trial design, the risk of contaminating intervention and control groups is high and the logistic and economic challenges of multicenter studies are formidable, especially without previous strong evidence of impact. Therefore, single-site impact analysis is important because it measures the actual effects of using the EDSS

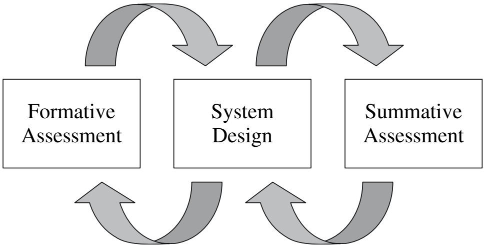


Fig. 5. The process of building an EDSS is an iterative cycle of forming the system around user needs, designing appropriate interactions between the system and users, and then evaluating the true impact of the system using quantitative studies ***(8)***.

in clinical practice, which is critical for planning of successful multi-site studies ***(4)***.

An important objective of EDSS evaluation is quantitative assessment of potential impact of EDSS on patient outcomes, work practices and the introduction of new errors. The potential benefits of EDSS can be summarized into three groups:

1. Improved patient safety
    a. Reduction of medical errors
    b. Enhancement of clinical decisions and resource utilization
2. Improved quality of care
    a. Improved compliance with guidelines and clinical protocols
    b. Improved access to and use of evidence
    c. Improvements in the patient satisfaction and the patient consent process
3. Improved efficiency of healthcare delivery
    a. Reductions in costs and in physician time spent on administrative tasks
4. Optimization of resource allocation because of:
    a. The individualized selection of procedure types and post-procedure follow-up
    b. Optimization of personalized therapeutic modalities based on individual molecular risk profiles
    c. Cross-disciplinary treatment paradigm

Outcome measures for DSS should include predictive values, as well as safety and efficiency. For clinicians, negative predictive value and safety are most important because their primary concern is to minimize “missed” patients who have the targeted outcomes. For insurers, positive predictive value and efficiency are the most important because their major concern is cost-effectiveness. Accuracy and other measures (sensitivity, specificity, and area under the receiver-operating characteristic curve (ROC)) may be misleading because they assume equivalent social value for true-positive and true-negative results and may vary with the overall prevalence of outcomes.

Current evidence on the impact of bioinformatics EDSS is limited. It has been documented that they can serve as an educational tool for low-risk patients or can be a useful adjunct to genetic counseling for those at high risk. For example, evidence from randomized controlled trials suggests that an interactive decision support is more effective than standard genetic counseling for increasing knowledge of breast cancer and genetic testing among women at low risk of carrying a mutation ***(13)***. The beneficial impact is more likely when an EDSS provides specific recommendations, or provides them automatically as part of clinicians' routine workflow. However, beneficial impact in a research study (efficacy) does not guarantee beneficial impact in clinical practice (effectiveness).

4. Conclusions

Successful decision support system design should be task-specific and address situational response requirements and environmental characteristics such as complexity and information overload. Electronic decision aids can reduce decision errors ***(4)*** and also enhance what has become the shared and collaborative process of the use of "omics" technologies for the diagnosis and management of diseases. The dichotomy between the proliferation of evidence such as clinical practice guidelines, and its low uptake, indicates that clinicians are already struggling with information over-supply and concomitant competition for their attention ***(44,49)***. This has lead to the suggestion that the notion of the "best evidence" should be replaced with a more complex notion of the "most effective evidence delivery," which takes into account both the inherent potential of evidence to improve clinical decisions, as well as the likelihood that its mode of delivery will be adopted ***(8)***.

There is a growing demand for tools to support the capture of genomic results as generated by molecular diagnostic and cytogenetic methods, appropriate controlled vocabularies, and applications enabling clinicians to utilize these results to support their decision-making. Success of EDSS in clinical bioinformatics will require planning robust prospective trials, analysis of health care outcome and economic data, and developing new healthcare delivery models. Indeed it is unlikely that the vision for personalized medicine will not be fully realized without workflow integrated, and genomics based, clinical decision support systems.

5. Notes

1. A straightforward electronic risk assessment tool for breast cancer developed by scientists at the US National Cancer Institute and the National Surgical Adjuvant Breast and Bowel Project allows a risk to be calculated for invasive breast cancer www.cancer.gov/bcrisktool/. However, this tool demonstrates some the complexities involved in electronic decision support. For example, the tool is not useful in difficult cases such as ones with a known *BRCA1* or *BRC2* mutation or cases with an earlier cancer or locular carcinoma in situ or ductal carcinoma in situ. One of the seven questions used to assess risk asks for the woman's race/ethnicity. The five ethnic groups given include: White, Black, Hispanic, Asian or Pacific and American Indian. However, responding to any of these groups except for

"White" will provoke a disclaimer indicating that data on non-White ethnicities are uncertain and so may not be accurate until more information is generated.

2. Data provided in reference ***(10)*** indicate that surgery in the USA costs $450 billion per year. On top of this there are additional costs related to complications which total $25 billion. The latter costs will only increase as more surgery is conducted on an increasingly ageing population. Pre-operative risk assessment tools to guide perioperative management of high-risk patients are available but their predictive value is very poor. Hence, a new and alternative approach is "perioperative genomics" which is being used to determine why patients respond so differently to a surgical intervention. The first step is to identify what genes might contribute to post-operative complications, e.g., genes for inflammation, thrombosis, cardiac arrhythmias, wound healing, infection, shock and so on. A genetic "fingerprint" of these genes is then obtained pre-operatively so that an individual's particular risks can be identified early, and appropriate preventative measures put into place. Getting this genetic profile will only the beginning. The assimilation of the results as well as their overall interpretation for the clinician will require informatics-based decision algorithms. A start along the approach described has already been made to predict graft rejection. It is called the AlloMap^{TM} in which the expression of 20 genes is measured by quantitative PCR and then translated into a clinically actionable score that can be used to diagnose cardiac allograft rejection early and non-invasively (10, www.allomap.com/). However, more sophisticated genomics and informatics will be required to predict those at risk of post-operative complications.

References

1. Molidor, R., Sturn, A., Maurer, M., and Trajanoski, Z. (2003) New trends in bioinformatics: from genome sequence to personalized medicine. *Exp. Gerontol.* **38**, 1031–1036.
2. Billings, P. R., Carlson, R. J., Carlson, J., Cain, M., Wilson, C., Shorett, C., et al. (2005) Ready for genomic medicine? Perspectives of health care decision-makers. *Arch. Intern. Med.* **165**, 1917–1919.
3. Osheroff, J. A., Teich, J. M., Middleton, B. F., Steen, E. B., Wright, A., and Detmer, D. E. (2006) A roadmap for National Action on Clinical Decision Support. American Medical Informatics Association, June 2006, p24 (see http://www.amia.org/inside/initiatives/cds/)
4. Reuna, V. F., Lloyd, F. J., and Whalen, P. (2001) Genetic testing and medical decision making. *Arch. Intern. Med.* **161**, 2406–2408.
5. Coulson, A. S., Glasspool, D. W., Fox, J., and Emery, J. (2001) RAG: A novel approach to computerized genetic risk assessment and decision support from pedigrees. *Method. Inf. Med.* **40**, 315–322.
6. Cruz, J. A., and Wishart, D. S. (2006) Application of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 59–78.

7. Liu, T. F., and Shafer, R. W. (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect. Dis.* **42**, 1608–1618.
8. Coiera E. (2003) *Guide to Health Informatics*. 2nd ed. Arnold, London.
9. Freedman, A. N., Seminara, D., Gail, M. H., Hartge, P., Colditz, G. A., Ballard-Barbash, R., et al. (2005) Cancer risk prediction models: a workshop on development, evaluation, and application. *J. Natl. Cancer Inst.* **97**, 715–723.
10. Podgoreanu, M. V., and Schwinn, D. A. (2005) New paradigms in cardiovascular medicine. Emerging technologies and practices: perioperative genomics. *J. Am. Coll. Cardiol.* **46**, 1965–1977.
11. Barrera, J., Cesar, R. M., Ferreira, J. E., and Gubitoso, M. D. (2004) An environment for knowledge discovery in biology. *Comput. Biol. Med.* **34**, 427–447.
12. Baxevanis, A. D., and Ouellette, B. F. F. (eds.) (2005) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. 3rd ed. John Wiley & Sons, England.
13. Green, M. J., Peterson, S. K., Baker, M. W., Harper, G. R., Friedman, L. C., Rubinstein, W. S., et al. (2004) Effect of a computer-based decision aid on knowledge, perceptions, and intentions about genetic testing for breast cancer susceptibility: a randomized controlled trial. *J. Am. Med. Assoc.* **292**, 442–452.
14. Mellmann, A., Friedrich, A. W., Rosenkotter, N., Rothganger, J., Karch, H., Reintjes, R., et al. (2006) Automated DNA sequence-based early warning system for the detection of methicillin-resistant *Staphylococcus aureus* outbreaks. *PLoS Med* **3**: e33.
15. Seo, D., and Ginsburg, G. S. (2005) Genomic medicine: bringing biomarkers to clinical medicine. *Curr. Opin. Chem. Biol.* **9**, 381–386.
16. Gerling, I. C., Solomon, S. S., and Bryer-Ash, M. (2003) Genomes, transcriptomes, and proteomes. *Arch. Intern. Med.* **163**, 190–198.
17. Sintchenko, V., and Coiera, E. (2003) Which clinical decisions benefit from automation? A task complexity approach. *Int. J. Med. Inform.* **70**, 309–316.
18. Wallace, C. S., and Patrick, J. D. (1993) Coding decision trees. *Machine Learn.* **11**, 7–22.
19. Chu, P. C., and Spires, E. E. (2000) The joint effect of effort and quality on decision strategy choice with computerised decision aids. *Dec. Sci.* **31**, 259–292.
20. Sitter, H., Prunte, H., and Lorenz, W. (1996) A new version of the program ALGO for clinical algorithms, in *Medical Informatics Europe 1996/Studies in Health Technology and Informatics* (Brender, J., Christensen, J. P., Scherrer, J. R., and McNair, P. eds.), IOS Press, Amsterdam, pp. 654–657.
21. Edwards, G., Kang, B. H., Preston, P., and Compton, P. (1995) Prudent expert systems with credentials: managing the expertise of decision support systems. *Int. J. Biomed. Comput.* **40**, 125–132.
22. Simon, R. (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J. Cancer* **89**, 1599–1604.
23. Green, B. T., and Khan, J. (2004) Diagnostic classification of cancer using DNA microarrays and artificial intelligence. *Ann. NY Acad. Sci.* **1020**, 49–66.

24. Kapetanovic, I. M., Rosenfeld, S., and Izmirlian, G. (2004) Overview of commonly used bioinformatics methods and their applications. *Ann. NY Acad. Sci.* **1020**, 10–21.
25. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2005) Machine learning in bioinformatics. *Brief Bioinform.* **7**, 86–112.
26. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
27. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U S A* **99**, 6567–6572.
28. Hedenfalk I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., et al. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**, 539–548.
29. Breiman, L. (1996) Bagging predictors. *Machine Learn.* **24**, 123–140.
30. Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Stat.* **26**, 1651–1686.
31. Neville, P., Tan, P., Mann, G., and Wolfinger, R. (2003) Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum. *Proteomics* **3**, 1710–1715.
32. Wagner, M., Nalik, D., and Pothen, A. (2003) Protocols for disease classification from mass spectrometry data. *Proteomics* **3**, 1692–1698.
33. Malyarenko, D. I., Cooke, W. E., Adam, B. -L., Malik, G., Chen, H., Tracy, E. R., et al. (2005) Enhancement of sensitivity and resolution of surface-enhanced laser description/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clin. Chem.* **51**, 65–74.
34. Jain, A. K., Duin, R. P. W., and Jianchang, M. (2000) Statistical pattern recognition: a review. *IEEE Trans. Pattern. Anal. Mach. Intell.* **22**, 4–37.
35. Shin, H., and Markey, M. K. (2006) A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *J. Biomed. Inform.* **39**, 227–248.
36. Guyon, I., and Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182.
37. Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. *Artif. Intell. Med.* **32**, 71–83.
38. Tatay, J. W., Feng, X., Sobczak, N., Jiang, H., Chen, C., Kirova, R., et al. (2003) Multiple approaches to data mining of proteomics data based on statistical and pattern classification methods. *Proteomics* **3**, 1704–1709.
39. Hilario, M., Kalousis, A., Muller, M., and Pellegrini, C. (2003) Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics* **3**, 1716–1719.

40. Witten, I. H., and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann Publishers, San Francisco, CA.
41. Gardner, S. P. (2005) Ontologies and semantic data integration. *Drug Discov. Today Biosilico* **10**, 1001–1007.
42. McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., et al. (2003) LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-year update. *Clin. Chem.* **49**, 624–633.
43. Cimino, J. J. (2000) From data to knowledge through concept-oriented terminologies: experience with the medical entities dictionary. *J. Am. Med. Inform. Assoc.* **7**, 288–297.
44. Ohno-Machado, L., Gennari, J. H., Murphy, S. N., Jain, N. L., Tu, S. W., Oliver, D. E., et al. (1998) The guideline interchange format: a model for representing guidelines. *J. Am. Med. Inform. Assoc.* **5**, 357–372.
45. Peleg, M., Boxwala, A. A., Bernstam, E., Tu, S., Greenes, R. A., and Shortliffe, E. H. (2001) Sharable representation of clinical guidelines in GLIF: relationship to the Arden syntax. *J. Biomed. Inform.* **34**, 3170–3181.
46. Eastwood, A., and Sheldon, T. (1996) Organisation of asthma care: what difference does it make? A systematic review of the literature. *Qual. Health Care* **5**, 134–143.
47. Smith, H. L., Bullers, W. I., and Piland, N. F. (2000) Does information technology make a difference in healthcare organization performance? A multiyear study. Hospital topics. *Res. Perspect. Health Care* **78**, 13–22.
48. Mant, D. (1999) Can randomized trials inform clinical decisions about individual patients? *Lancet* **353**, 743–746.
49. Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P.A., et al. (1999) Why don't physicians follow clinical practice guidelines? A framework for improvement. *J. Am. Med. Assoc*. **282**, 1458–1465.
50. Berg, M. (2001) Implementing information systems in health care organizations: myths and challenges. *Int. J. Med. Inform.* **64**, 143–156.
51. Eisenberg, J. M. (1999) Ten lessons for evidence-based technology assessment. *J. Am. Med. Assoc*. **282**, 1865–1869.
52. Slotnick, H. B. (2000) Physicians' learning strategies. *Chest* **118**, 18–23.
53. McAlister, F. A., Graham, I., Karr, G. W., and Laupacis, A. (1999) Evidence-based medicine and the practicing clinician. *J. Gen. Intern. Med.* **14**, 236–242.
54. Reilly, B. M., and Evans, A. T. (2006) Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann. Intern. Med.* **144**, 201–209.

19

eConsulting

Siaw-Teng Liaw and Peter Schattner

Summary

*e*Consulting, in all its contexts, can promote and improve the amount and quality of services and knowledge transferred to and among the community of health care providers and consumers. It can also improve the efficiency and effectiveness of the specialist and generalist workforce and accessibility to the services provided. This chapter defines *e*Consulting, provides the context, and introduces a conceptual framework to describe its current practice and future possibilities. A clinical scenario of a patient with a breast lump is used to ground the molecular, clinical, organizational, and social, legal, and ethical issues in real world practice. The approach/method used is based on the clinical process, evidence-based practice, and appraising the quality, validity, relevance, and usefulness of the information. The practicalities and utility of current *e*Consulting tools are discussed with a view to future ubiquitous use. Working through this chapter should assist readers to understand and describe (1) how *e*Consultations can link and translate scientific research into clinical practice, (2) the current implications of *e*Consultations, (3) the future potential of *e*Consultations.

Key Words: Internet, *e*Health, *e*Consultations, evidence-based practice, quality of online information, online health advisors, usefulness, validity, relevance.

Abbreviations: ANT – actor network theory; EDS – electronic decision support; GP; – general practitioner (family physician/doctor, primary care physician, primary care practitioner); ICT – information and communications technology; HON (code) – Health on the Net (code)

From: *Methods in Molecular Medicine, Vol. 141: Clinical Bioinformatics*
Edited by: R. J. A. Trent © Humana Press, Totowa, NJ

1. Introduction

1.1. The Context for eConsulting

The broadest context for *e*Consulting is *e*Commerce, which covers everything we do as a society. This chapter focuses on *e*Health, *e*Learning and *e*Research, which addresses the applications within an academic community to facilitate the research, creation and sharing of knowledge to improve clinical practice and health outcomes. *e*Health involves the combined use of electronic information and communication technology (ICT) to transmit, store and retrieve digital data electronically for health and administrative purposes, locally and at a distance. It can improve access to health services and information by all citizens.

*e*Health services include electronic health records and information networks, telehealth and online services, personal and portable communication systems, health portals, and decision support tools to assist prevention, diagnosis, treatment, health monitoring, and personal lifestyle management. The ubiquity of access and low cost of distribution of information using ICT suggest that *e*Health services can promote cost-efficiencies, facilitate coordination and integration in the health system, and improve equity of access to health services, education and information by patients, healthcare professionals, healthcare managers and authorities ***(1)***. *e*Health can underpin and strengthen the organizational relationships and behaviour within the health care team as a focus for inter-sectoral and inter-professional service integration, research and knowledge transfer among universities, health services and the community.

The Royal Flying Doctor Service ***(2)***, the first *e*Health project in Australia, is a precursor of more recent *e*Health projects in this community ***(3***,***4)***. Likewise, the Australian School of the Air has evolved to more sophisticated *e*Learning programs such as Rural and Remote Medical Education Online (www.rrmeo.org.au) and the Royal Australian College of General Practitioners Online (www.racgp.org.au). *e*Learning and knowledge transfer can improve the therapeutic and educational relationships and interactions among all the "actors" in the network where health care and health education occur: health care providers, consumers, managers, administrators, scientists/researchers, educators, legislators and policy makers. The generic information sharing process in this actor network ***(5)*** is the "consultation".

1.2. eConsulting Defined

According to the Oxford English Dictionary, a consultation involves two or more parties taking counsel, conferring about, deliberating upon, considering, meditating, planning, and contriving. The exchange of knowledge, information and data among the "actors" (people and systems) in a consultation is the information flow. *e*Consulting is the use of electronic data transfer in cyberspace (Internet) for information exchange and financial transactions, using voice, video, and/or data transmission systems, between two or more actors at two or more sites. The *e*Consultation may be

1. Synchronous, where the "actors" are present and interacting at the same time, e.g., teleconferencing
2. Asynchronous, where the "actors" do not have to be present and interacting at the same time, e.g., discussion forums as well as more sophisticated programs like knowledge-based *online health advisors*.

1.3. Models of eConsulting

There are two models for *e*Consultations:

1. Consulting model: a style of interaction in which an electronic decision support (EDS) program serves as an adviser, accepting patient-specific data, asking questions, and generating advice for the user about diagnosis and management. A consulting system develops and suggests problem-specific recommendations based on user input.
2. Critiquing model: a style of interaction in which an EDS program serves as a sounding board for the user's ideas, expressing agreement or suggesting reasoned alternatives. A critiquing system evaluates and suggests modifications for plans or data analyses already formed by a user.

A combination of consulting and critiquing approaches is usual in practice.

1.4. Tools for eConsulting

*e*Consulting tools, which can be web-based or desktop applications, may focus on communication, conferencing or collaborative management or co-ordination of activities or resources. *Electronic communication tools* send messages, files, data, or documents between "actors", facilitating information sharing. Examples include fax, e-mail, listservs, voice mail and web publishing. *Electronic conferencing tools* such as audio/video conferencing, Internet forums or chat rooms facilitate the interactive sharing of information and real-time text messages. More sophisticated examples are data conferencing using networked PCs with a common "whiteboard," and electronic meeting systems (EMS) in rooms equipped with a screen projector and networked PCs. *Collaborative management tools (or groupware)* include time management software to schedule and remind group members about events; project management systems to schedule, track, and chart group tasks and activities; workflow systems to manage tasks and documents within a business process; knowledge management systems to collect, organize, manage, and share various forms of information such as "*online health advisors*" ***(6)***; and social software systems to organize social relations of groups of "actors" (www.darwinmag.com/read/050103/social.html).

1.5. "Actors" in eConsulting

The Actor-Network Theory (ANT) is a widely used approach to describe and explain the complex relationships and interactions in human communities and health service organizations ***(5)***. Developed originally to study scientific practices ***(7)***, the ANT has become a generic framework to understand social phenomena in health services and systems research. ANT defines society as networks of heterogeneous actors, both human and non-human. It maps relations that are simultaneously material (*between things*) and "semiotic" (*between concepts*), e.g., the personal and professional interactions in a bank or laboratory involve both people and their ideas, their equipment and their computers – together, these form a single network. Society, the economy, organizations, families, agents, computing systems and machines are all effects generated through the material-semiotic interactions of actor-networks ***(5)***. A person is not an isolated entity but is always linked to a heterogeneous network of resources and agents that define him or her. Such actor-networks are not intrinsically coherent and may contain conflicts, e.g. poor labor relations or incompatible computer software.

Consider a human genetics agency which provides advisory, support and training services to the community. The straightforward aspects of the service can be provided directly to the consumer as a *consumer-oriented online genomics health advisor*, with support by an automated frequently-asked-question (FAQ) set-up and a "help desk". *Provider-oriented online resources* can support trained clinicians to conduct more complex genomic consultations and specialists and genomic counselors who support them. The system can support the actors in the *e*Health and *e*Learning actor-network. The ANT approach can facilitate a more actor-centric clarification of the extent to which the agency interacts with, shapes and is shaped by people, other technologies, and institutions. This enables more actor-specific technology assessment, budgets, professional development, consumer engagement and rigorous public debate and analysis about the social, ethical and policy implications of genomic knowledge and services ***(8)***. A possible model is the US National Institutes of Health-sponsored **GeneTests** Web site (http://genetests.org/), a free medical genetics information resource developed for clinicians, teachers and researchers.

The "actors" in an *e*Consultation can be the community, government, organizations, clinicians, academics, managers, and software applications (**Table 1**).

Table 1
An actor framework for *e*Consulting

Actor (consultant)	Action	Objective/Outcome	Actor(consultee)

Community (consumer & patient)	Source of personal information;
	seek & take counsel;
	give counsel
		Make appropriate and informed decisions about personal health;
		appropriate and optimal use of health services
			Health care provider;
			health information provider;
			health services;
			government

Clinician (doctor, nurse, allied health professional)	Source of formal knowledge;
	seek & take counsel;
	give counsel;
	consults / uses information
		Make evidence – based clinical decisions;
		deliver safe, efficient and effective health care
			Patients & care-givers;
			clinicians;
			families & communities;
			health services;
			health information services;
			government

Academic (researcher & teacher)	Source of knowledge;
	seek & take counsel;
	consults / uses information
		Develop and deliver health information to support evidence – based practice and safe, efficient and effective health care and services	Patients & care-givers;
			clinicians;
			families & communities;
			health services;
			health information services;
			government

Manager & planner	Source of knowledge;
	seek & take counsel;
	give counsel
		Develop and deliver safe, efficient, and effective health services	Patients & care-givers;
			clinicians;families & communities;
			health services;
			health information services;
			government

(Continued)

Table 1
(Continued)

Actor (consultant)	Action	Objective/Outcome	Actor (consultee)

Software application, e.g., electronic decision support or online health advisor systems	Consulting model:
	■ Consults and gives professional counsel
	Critiquing model:
	■ Facilitates reflection and provides affirmation and / or alternatives
		Support development and delivery of
		■ Safe, efficient and effective health services;
		■ Evidence-based health information;
		by clinician, manager, planner & policymaker
			Patients & care-givers;
			clinicians;
			families & communities;
			health services;
			health information services;
			government

Professional organizations	Source of formal knowledge;
	consult populace about professional issues;
	provide professional information to government and public
		Develop evidence-based health information, policy, standards & benchmarks;
		improve public's health and knowledge
			Patients & care-givers;
			clinicians;
			families & communities;
			health services;
			health information services;
			professional organisations;
			governments and agencies

Government(all levels & their agencies)	Source of formal knowledge;
	consult populace about relevant issues;
	provide health and policy information to professional bodies and public
		Develop evidence-based health information, policy, standards & benchmarks;
		improve public's health and knowledge
			Patients & care-givers;
			clinicians;
			families & communities;
			health services;
			health information services;
			professional organisations;
			governments and agencies

While the most widely documented form of *e*Consulting is online consulting by government and in *e*Commerce (www.wikipedia.org), our focus will be clinical *e*Consulting as represented in the following scenario.

2. Scenario: Sally has Breast Cancer

The process and impacts of *e*Consultations will be described and demonstrated in the clinical context of Sally, a patient with a breast lump:

> Sally, 50 years old, presents with a small breast lump. The family doctor (GP) who examines her is uncertain if it is benign or malignant, and suggests a mammogram as the next step. The patient questions whether she should have a mammogram as there is radiation involved. If the doctor is not sure whether there is a lump or not, shouldn't she get a specialist opinion first?
>
> The GP, recognizing that a normal mammogram in the presence of a breast lump is almost invariably an indication for a biopsy, refers her to the local breast cancer specialist. The appointment is made using email. The specialist's examination and subsequent mammogram and biopsy confirm breast cancer.
>
> The management choices are surgery, radiotherapy, and chemotherapy or hormone manipulation, depending on the stage of the cancer. A diagnosis of stage 2 breast cancer (tumor is more than 2 cm but less than 5 cm) is made and Sally starts on radiation with adjuvant chemotherapy (*see* **Note 1**).
>
> Martha is Sally's first cousin on her father's side. She is aged 57 years and has been taking hormone replacement therapy for the past 4 years. She consults the GP as she is concerned about the risk of breast cancer. She has gone to her computer at home and "Googled" "breast cancer and screening," which has added to her confusion. How should the GP advise Martha?
>
> Sally's mother also consults the GP. She was born in Poland and emigrated from there via Israel in the 1950s. She had been a smoker and had used some form of oral contraception and, perhaps, hormone replacement therapy as well. What advice should the GP give Sally's mother who is particularly worried how this family susceptibility might affect her family in their environment and workplace.
>
> Sally has nearly completed her therapy and her family is planning a holiday in Bali to celebrate. She seeks travel advice. How would her chemotherapy be affected by any immunisations, e.g., Hepatitis A? How should she deal with "travelers' diarrhea"? Does Sally's family history of breast cancer means she is susceptible to other cancers or modalities of treatment? The GP looks for information related to these questions on the Internet.

This typical clinical scenario indicates that *e*Consulting, defined as consulting involving the electronic transfer of data, is not routine clinical practice. The tools are available and occasionally used as demonstrated by appointments being made by email, results and referral documents communicated via the Internet, and *online health advisors'* being used for travel medicine.

Based on the above scenario, the following aspects of *e*Consulting will be described: evidence-based practice in screening, diagnosis, management, monitoring and review, social, workplace, ethical and legal issues.

2.1. Clinician's Use of Online Health Information in eConsultations

Searching the Cochrane Collaboration Database of Systematic Reviews for screening and management options resulted in a number of reviews by the Cochrane Breast Cancer Group. The focus was specific treatments. Nevertheless, a review on breast screening ***(9)*** and follow-up protocols ***(10)*** caught the clinician's interest. Medline was also searched for breast screening but the list was too long to review during the consultation. Two promising references were flagged for further reading: (1) Kriege M, Brekelmans C, Obdeijn I, et al. Factors Affecting Sensitivity and Specificity of Screening Mammography and MRI in Women with an Inherited Risk for Breast Cancer. Breast Cancer Res Treat 2006. (2) Advisory Committee on Breast Cancer Screening. Screening for breast cancer in England: past and future. J Med Screen 2006;13:59–61. The clinician also undertook to examine the UK NHS and OncoLink sites later. For more generic genetic information, the clinician also found the Australasian Genetics Resource Book at www.genetics.com.au.

2.2. Consumers Use of Online Health Information in eConsultations

Martha's "Googling" resulted in a range of hits of variable quality and purposes, which added to her level of confusion (**Table 2**). This experience is consistent with that reported by the 20% of the population who use the Internet for online health information ***(11)***.

The ordered list of "hits" was (1) "breastcancer.org," a HON endorsed non-profit organization dedicated to providing information and community to those affected by breast cancer; (2) a site, with a similar name, advertising Faslodex (fulvestrant), a hormonal treatment for hormone receptor-positive metastatic breast cancer in postmenopausal women; (3) a news site about breast cancer screening; (4) a breast cancer genetic screening site hosted by the Lawrence Berkeley National Laboratory; (5 and 6) the comprehensive UK NHS Breast Screening site with the latest research information about the NHS Breast Screening Programme; (7 and 8) the comprehensive US National Cancer Institute site; (9) the American Cancer Society, which provided good explanations of the change in breast screening guidelines; (10) the U.S. Preventive Services Task Force guidelines; (11) Oncolink, a product of the University of Pennsylvania Abramson Cancer Centre, which also had some very

Table 2
Martha's search on Google produced the following hits on the first page (2006 June 25)

Breast Cancer Screening www.**breastcancer**.org

Current research on detection tests with plain-English explanations.

Breast Cancer Screening www.seek-your-options.info

Advanced **Breast Cancer** Treatment Visit Here for Treatment Options

News results for **Breast cancer screening** - View today's top stories

Ethnic women more likely to shun **cancer screening** - Globe and Mail - 23 Jun 2006

Footing bill for **cancer** scanner - Scotsman - 23 Jun 2006

Nigeria: **Breast Cancer** Not a Killer Disease, Says Expert - AllAfrica.com - 20 Jun 2006

Breast Cancer Screeningwww.lbl.gov/Education/ELSI/**screening**-main.html - 4k - 22 Jun 2006

Breast Cancer and Genetic **Screening**. **Breast Cancer** Awareness Emblem **...**

Some of the most recent discoveries in this field involve **breast cancer**. **...**

NHS **Breast Screening** Programme (NHSBSP)

www.**cancerscreening**.nhs.uk/**breast**screen/

The NHS **Breast Screening** Programme Home Page provides information about the **screening** programme to both National Health Service personnel and the public.

NHS **Cancer Screening** Programmes www.**cancerscreening**.nhs.uk/ - 10k

NHS **Cancer Screening** Programmes: **Breast Screening**, Cervical **Screening**,

Bowel **Cancer Screening** and Prostate **Cancer** Risk Management.

(Continued)

Table 2
(Continued)

Breast Cancer: **Screening** and Testing - National **Cancer** Institute
Methods of **cancer** detection, including information about new imaging technologies, tumor markers, and biopsy procedures.
www.**cancer**.gov/**cancer**topics/**screening/breast** - 29k **Breast Cancer Screening** - National **Cancer** Institute
www.**cancer**.gov/**cancer**topics/pdq/**screening/breast/**patientExpert-reviewed
information summary about tests used to detect or screen for **breast**
cancer.ACS :: Updated **Breast Cancer Screening** Guidelines Released
www.**cancer**.org/docroot/NWS/content/NWS_1_1x_Updated_Breast_Cancer_Screening_Guidelines_Released.asp - 36k
Breast Cancer: **ScreeningScreening** for **Breast Cancer**. Release Date: February 2002. Summary of Recommendations /
Supporting Documents. Summary of Recommendations **...** www.ahrq.gov/clinic/uspstf/uspsbrca.htm - 7k - 23 Jun 2006
Breast Cancer Information and Resources | Oncolink**Breast Cancer** information including risk, prevention, **screening**,
symptoms, research, treatment, and support. Provided by Oncolink - The Web's First **Cancer ...**
www.oncolink.org/types/article.cfm?c=3&s=5&ss=33&id=8320 - 44k - 23 Jun 2006 **Cancer** - National **Breast** and Cervical
Cancer Early Detection **...**CDC provides access to critical **breast** and cervical **cancer screening** services for underserved
women in the United States, the District of Columbia, **...**
www.cdc.gov/**cancer**/nbccedp/ - 54k **Screening** Mammography for **Breast Cancer**The effectiveness of **breast cancer**
screening by mammography in younger women. **... Breast cancer screening** with mammography: overview of Swedish
randomized **...** acpm.org/**breast**.htm - 13k

comprehensive information for patients and clinicians; (12) the CDC Cancer Prevention and Control Program, which "provides access to critical breast and cervical cancer screening services for underserved women in the United States, the District of Columbia, 4 U.S. territories, and 13 American Indian/Alaska Native organizations"; and (13) the American College of Preventive Medicine Practice Policy Statement in Screening Mammography, a comprehensive and referenced paper.

The Google algorithm for ranking the sites by the number of links allowed a commercial site to come out with the search term "breast cancer screening." That only one site had HON endorsement warrants further examination (*see* **section 3.3** below for a more comprehensive description of the HON code). The Wikipedia had a very comprehensive entry for "breast cancer screening," with a range of credible links. Consumers use online health information to find out more about the cause or description of their disease, as a second opinion, or to enable them to discuss it with their doctor or pharmacist, or change their health care management. General search engines like Google are probably just as good as specialized ones when searching for health information for consumers ***(12)***. While the use is increasing, we know very little about any health effects of consumer use of online health information ***(13)***

2.3. Screening and Diagnostic DNA Tests

In assessing online information, Sally needs to be aware of the differences between diagnostic and screening tests. We *test* an individual to diagnose a condition that other evidence suggests may be present. In contrast, we *screen* all members or groups of a community or family for a condition where there is no prior evidence of its presence in the individual. The "odds" of a positive test when someone has the disease are different in the two situations. In Sally's case, a diagnosis was made by examining the anatomy of her breasts with a mammogram, and the histology of her "lump" with direct microscopy. Sally's relatives are also screened with mammograms to try to detect breast lumps early as well as with genetic (DNA) tests to detect the presence of mutations in *BRCA1* or *BRCA2*, which indicates genetic susceptibility to breast (and ovarian) cancer before the onset of any symptoms. If there is genetic susceptibility, the evidence suggests that Magnetic Resonance Imaging (MRI) is a better screening modality than mammography ***(14)***.

In general, genetic tests are done to

1. Screen for common hereditary disorders in the population, e.g., phenylketonuria
2. Screen for genetic abnormalities in people at risk, e.g., relatives of people with hereditary disorders and certain ethnic groups

3. Assist the diagnosis for difficult-to-diagnose illnesses
4. Identify paternity and other important relationships when uncertain.

Insurance companies claim genetic testing could add to the array of health and lifestyle information they use to set premiums. However, opinion is divided as to whether customers should be encouraged or required (by regulation or policy) to be genetically tested. This is important in an environment where home testing kits for genetic predispositions are becoming increasingly available in the market place. The social, economic, legislative, technical and behavioral complexity underpinning the current status of genetic screening can be explained using Actor-Network Theory with the genetic test as an active participant (an actor) in the socio-technical network. It clarifies the extent to which the test interacts with, shapes and is shaped by people, other technologies, businesses, institutions and government ***(15)***. In Sally's case, her fears about cancer and radiation influenced the clinician's decision, along with his relationships and interactions with other actors such as the imaging and breast cancer specialists, Sally's family network, medico-legal conditions, defensive practice, and so on. The material-semiotic approach will invoke an understanding of Sally and her doctor's confidence in the evidence as well as the quality of the equipment and the practitioners.

2.4. Pharmacogenetics to Guide Drug Treatment

To support Sally in her quest for information about the best treatments for breast cancer, clinicians, clinical geneticists, and genetic counselors should have a functional knowledge of the rapidly emerging field of pharmacogenetics (it is also called pharmacogenomics). This is the study of the variability in the information in a gene and its gene product observed in a population that is associated with susceptibility to develop a condition, as well as determining drug response. Pharmacogenetics looks for genetic differences within a population that explain certain observed responses to a drug or susceptibility to a health problem. Pharmacogenetic-based studies are rapidly elucidating the inherited nature of these differences in drug disposition and effects, thereby enhancing drug discovery and providing a stronger scientific basis for optimizing drug therapy on the basis of each patient's genetic constitution ***(16–18)***. Herceptin may be useful for Sally if her *HER2* (human epidermal growth factor receptor-2) gene is being over-expressed ***(19)***. In the context of drug design and clinical research and development, a measurable and unequivocal definition of a phenotype, such as drug efficacy or toxicity, is essential. The mechanisms of phenotypic expression has evolved to include "modifier genes," such as those involved with drug metabolizing enzymes, their receptors and drug transporter genes, when describing disorders as diverse as risk of cancer, bone marrow toxicity resulting from occupational exposure, and Parkinson's disease.

A current challenge is to elucidate the multi-gene determinants of drug response. Genetic polymorphisms in drug-metabolizing enzymes, transporters, receptors, and other drug targets have been linked to inter-individual differences in the efficacy and toxicity of many medications. A further level of complexity is added by the increasing admixture amongst most ethnic groups; it is important to include ethnic information about DNA samples in all molecular epidemiologic studies. For instance, the clinician will want to explore if Sally's mother is Jewish.

2.5. Online Health Advisors

Online health advisors can provide specialist advice and support to generalist clinicians; the same could be said for consumers as in Sally's case. The US NIH-sponsored GeneTests Web site is an online health advisor with information about genetics services available. Sally would like a "second opinion" about the advice she has received. She visited the GeneTests website, located a genetic counselor and sought a second opinion for her advice as well as the role of genetic counseling for her cousin Martha.

The clinician replied that in these circumstances Martha would benefit from genetic counseling. However, because of a lack of genetic counselors in the region, arrangements were made for Martha to consult with one from the city teaching hospital, using videoconferencing.

2.6. Clinical Monitoring and Review

Clinical monitoring systems can be set up with remote clinician feedback loops. During her chemotherapy, Sally's white blood cell count was monitored and her GP automatically informed when it fell outside a predetermined range. This is analogous to a similar program with blood glucose monitoring that Sally's mother is using. Other remote monitoring systems, such as blood pressure measuring devices, can also be linked with web-based applications that can help patients care for their own conditions, with clinician guidance. This is particularly useful for chronic disease management in which patient self-care is of the utmost importance.

2.7. Security, Privacy, Confidentiality, and Medico-Legal Issues

*e*Consulting has introduced a new set of concerns about the privacy and security of health information. With genetic information, there are two other dimensions to consider: the family dimension and the likelihood that a genotype may express itself as a “disease”. This expands the actor-network to include the employer, insurance industry and the extended family, among others. Like most members of the public, Sally perceives that making her information electronic would make her information less secure and her life less private. The reality is that electronic records are more secure and private than paper-based ones. However, the increased access makes the loss of privacy more significant should security breaches occur. Anybody in the actor-network can be involved should a hacker or an employee circulate her health information intentionally or unintentionally, e.g. “reply all” option. There are technical safeguards to minimize these risks, including the use of passwords and encryption. Unfortunately, many clinicians view these solutions as cumbersome, and their use seems to be resisted at present.

An issue related to privacy and confidentiality is the use of Sally’s health information for other than clinical care e.g. research, quality assurance or professional development. Most countries have legislation and processes to cover information privacy. Examples of these include the USA Health Insurance Portability and Accountability Act of 1996 (http://www.hhs.gov/ocr/hipaa/), the Australian Privacy Act and Privacy Commissioner (http://www.privacy.gov.au/) and the UK Data Protection Act and Information Commissioner (http://www.ico.gov.uk/).

Other medico-legal issues can arise when doctors fail to respond to patients’ electronic questions in either a timely or clinically appropriate manner. For example, current clinical software for GPs allows pathology results to be downloaded into a holding file where they can be viewed by the GP and comments can be typed in for someone else to act upon. A practice nurse might be given the task of responding to patient phone calls about test results, and will be expected to relay the information provided by the GP. But what if the nurse’s interpretation of brief written advice is not completely correct? To what extent is the nurse, the GP or even the patient responsible for misunderstandings that can arise in these circumstances?

Another untested area concerns the quality and safety of electronic decision support, health advisors and other knowledge-based *e*Consulting tools available within desktop clinical software or online ***(20)***. How should they be regulated to ensure quality and safety? Should the clinical software or the knowledge-base sponsor or the clinician be responsible when errors of omission or commission occur? Who should be responsible for the accuracy and currency of the pharmacogenomic knowledge base underpinning a clinical information system? Who pays when *e*Consulting goes awry?

3. Methods: Evidence-based *e*Consulting

3.1. A Process for Evidence-based eConsulting

Figure 1 highlights the evidence-based clinical practice process and the information required to assist the decision-making by the various actors at different phases of the consultation. The consultation and *e*Consultation process can be divided into three broad phases:

1. To make an informed diagnostic decision, the clinician needs to know the validity, reliability and likelihood ratio of a symptom, sign, or diagnostic test
2. For management decisions, the clinician needs the secondary literature to inform him/her about the numbers needed to treat, numbers needed to harm and optimum follow-up, monitoring, and evaluation protocols
3. The clinician then reflects on his/her clinical practice, with input from the patient, to formulate further clinical questions specific to a clinical problem, the patient or the population s/he deals with. The reflection and its outcomes then feed back into the clinical practice process, starting the cycle again

All decisions will use reasonable patient preferences as a core factor for choices.

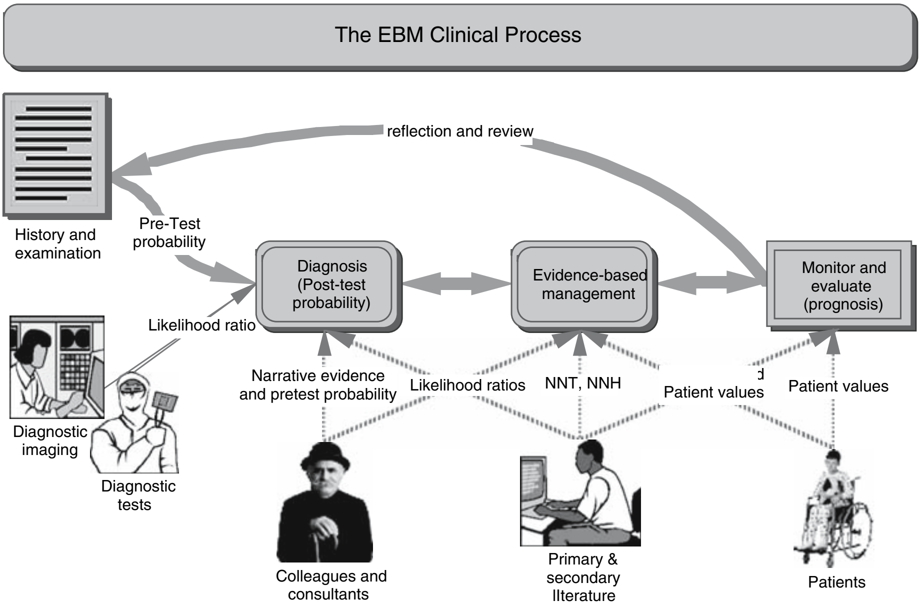


Fig. 1. The evidence-based clinical consultation. (Copyright ST Liaw)

3.2. Assessing Usefulness, Validity and Relevance

Online information must be presented such that it will be used in clinical practice. The evidence-based clinical consultation must be time efficient, i.e. quick and rewarding in terms of answers to the clinical questions at point of care. Slawson et al. *(21)* have conceptualized a "usefulness" equation for the answers to clinical questions (**Fig. 2**).

$$\textbf{Usefulness} = \frac{\textbf{Relevance x Validity}}{\textbf{Work i.e. effort required}}$$

Fig. 2. "Quantitating" usefulness (adapted from Slawson et al.)

Secondary sources of evidence, e.g., the Cochrane Database of Systematic Reviews, include structured reviews by experts in critical appraisal of primary evidence as found in original bibliographic citations, e.g., in Medline. Structured reviews, e.g., Critically Appraised Topics *(22,23)* and Patient-Oriented Evidence that Matters *(24)* are particularly useful for busy clinicians as they have been appraised for relevance and rigor by experts. Not all clinicians need to appraise primary evidence, but all clinicians need some skills in critical appraisal, especially of the secondary sources of evidence *(25)*. Knowing what is best practice is not enough—the best practice must be relevant, practical and sustainable! Searching for "evidence" can be undertaken by either the clinician or the patient, as explained earlier (*see* **2.1, 2.2**).

3.3. Assessing Quality of Information in eConsulting

The quality of online information is usually assessed according to consensus criteria for credibility, content, links, design and interactivity. The credibility criteria include source of information, disclosure of biases and errors, currency of information, relevance and utility of the information, and the credibility of the editorial review process. The content must be accurate, be transparent about the quality of evidence, state the original source of information, have an appropriate disclaimer, be logically organized with user-friendly navigation, and recognize the omissions. There should be back linkages, mechanisms for feedback, and an internal search engine. The quality of the links selected, in terms of the architecture and content, should be explicit and objective. The design is important. The interactivity should include comment options, chat rooms, and user-profiling.

While the rating of websites is highly variable and difficult at present ***(26)***, an approach for assessing the quality of health websites is the *Health on the Net Code of Conduct* (**HONcode**) (www.hon.ch/Conduct.html). This is a self-policing approach by groups that wish to abide by the HONcode principles; conforming organisations can display the HONcode logo on their site ***(27)***. The code's principles are as follows: (1) Medical advice is provided by qualified professionals. (2) The site supports (not replaces) physician-visitor relationship. (3) Confidentiality is respected. (4) Information is referenced to source data. 5 Claims are supported by evidence. (6) Information is provided in the clearest manner. (7) e-mail support is available. ***(8)*** Caveats are made explicit.

If the principles are violated and not corrected when requested, the HONcode symbol is removed from the website. There is debate about the effectiveness of the HONcode ***(28)***. There is some evidence that this coding is associated with accuracy of websites, although possible exceptions have been raised ***(29)***, including questions about fever in children ***(30)*** and patient asthma education ***(31)*** where the accessibility and quality are variable and the information needs of patients are often not met.

4. Encouraging *e*Consultation into the Future

How can we encourage eConsulting into the future? The enablers of and barriers to *e*Consulting may be personal, organizational, knowledge-related, or systemic/environmental. At the government and policy level, a proactive national *e*Health policy is essential. However, many countries, especially those without a nationalized health service like Australia and the United States, lack an explicit national *e*Health policy and an adequately funded implementation plan. There are no recommended benchmarks for bandwidth and standards to enable cost-efficient and effective sharing of information. The costs and affordability of ICT also varies across the country. The overall result is inadequate and inconsistent *e*Health infrastructure. Along with the lack of coordination and cooperation between the federal and state/provincial governments, there is often unsustainable duplication of activities like online information resources and telehealth programs. In Australia, there are more than 360 *e*Health projects, many of which are "…fragmented and uncoordinated, leading to problems of accessibility, scalability, duplication and lack of integration with existing systems" ***(3)***.

Confidence in *e*Health, its impact on work processes, lack of skills in using EDS, and concerns about medico-legal issues have been recognized as key factors influencing user acceptance and adoption of EDS ***(3)***. Clinicians will use *e*Consulting tools and applications if they are useful and relevant to their professional practice, easily incorporated into workflow and improve patient care, such as prescribing systems and decision support such as drug-drug interaction prompts ***(33,34)***. It is expected that the knowledge-base and information exchanged is evidence-based, valid and reliable.

The organization, e.g. hospital or general practice, must provide the supportive environment and tools for prompt and timely *e*Consulting. The effective use of *e*Consulting tools requires new and different skills, and the confidence to use them at point of care, require skilled and reliable technical training and professional support programs. The direct and indirect costs, especially in terms of investment of time and money, should also be supported by government as the ultimate outcome of *e*Consulting is to promote evidence-based practice and improve the safety and quality of care.

Consumers should also be similarly encouraged and supported to use *e*Consulting applications like "online health advisors" and other self-help applications. These tools should enhance the patient-clinician relationship, not detract from it. Issues associated with privacy, security and duty of care across state and national boundaries must be addressed. These issues are increasingly important as more sensitive genetic information become available and used in health care. User friendliness, accessibility and affordability of *e*Health are especially important in order to avoid further marginalizing existing socio-economically disadvantaged groups with a digital divide (*see* **Note 2**).

To promote the use of *e*Consulting, the strategy must be multi-pronged and targeted at government, organizations, learning institutions and the community.

5. Notes

1. Staging of Breast Cancer: Stage I—The cancer is no wider than 2 cm (about 1 inch) and has not spread outside the breast. Stage II—The tumor is more than 2 cm but less than 5 cm in the greatest dimension. Stage III —Tumor is more than 5cm in the greatest dimension. Stage IV—Tumor of any size with growth extending to the chest wall or skin.
2. How might Sally be managed in the future? The clinician begins his day with a cup of coffee pre-ordered by his computerized ordering system. He settles at his desk and checks his emails and results from requests done the previous day. Among the 50 or more emails, there is an appointment for Sally's session with the Clinical Genetics Advisory Service in the nearby capital city. Her genetic test results show a mutation in her *BRCA1* gene and the clinician is advised to counsel Sally, using the Clinical Genetics Advisory Service website as a support tool. Sally and the clinician explore the website and download information about the genetic risks of breast cancer. She also requests a videoconferencing session with her GP and oncologist, to discuss

further treatment, having first checked with www.quackwatch.com about alternative therapies for which there is no evidence of efficacy.

At a later routine follow-up consultation, the oncologist orders a telemammogram to check for new cancers. The radiological image is uploaded to Sally's Internet-based electronic health record. There is some calcium speckling on one of the films, and the radiologist and oncologist use radiological decision support software to help decide on the probability of the pattern signifying further neoplasia. The expert opinion, backed by calculations of electronic image patterns, suggests that the mammographic appearance is benign.

Sally continues with chemotherapy for her initial breast cancer. The drugs regimen used are being guided by the levels of the liver enzyme produced by the *CYP2D6* gene, which controls the metabolism of codeine to morphine ***(32)*** as well as the variant of the gene associated with slower caffeine metabolism (*1F). For monitoring, Sally tests her blood at home using a device that only requires a drop of blood. The blood is analyzed by a program configured to alert the oncologist if certain parameters are outside a predetermined range. Sally suffers mild nausea and notifies her oncologist about this side effect of her chemotherapy. He prescribes an anti-nausea drug in the electronic health record, which automatically notifies the nearest Internet pharmacy. The medication arrives by courier 6 hours later and Sally's nausea settles. Sally remains in good health under long-term review by her GP and oncologist.

Sally realizes that she still has some concerns about some unusual breast symptoms on the previously healthy side. She has some discomfort, and in spite of a negative telemammogram, she wishes to make further enquiries. Sally sends an email to the specialist asking about the likelihood of a mammogram missing a cancer when there is no lump but there is some localised pain and tenderness. She wonders whether the area should be biopsied "just to make sure." The oncologist believes that Sally has become overanxious, and refers her to a breast cancer listserv so that Sally can discuss her feelings with others who have been in a similar situation. Sally is not reassured and insists on a biopsy. Images of the histopathology slides are sent electronically to the foremost breast cancer unit in Washington. Finally, Sally is reassured, and from then on, continues in good health.

Sally joins an online self-help discussion group, the Breast Cancer Online Community, to discuss with other breast cancer sufferers how to manage their issues, including the handling of sensitive genetic information with members of the family. This self-help group is linked to a government-sponsored site which regularly consults the community about issues related to cancer and cancer genetics. Sally has contributed to both groups and feels that she may be able to mobilize her group to influence public opinion and government policy on genetic screening for breast cancer. She begins by doing online research.

References

1. Liaw, S. T., and Humphreys, J. S. (2006) Rural eHealth paradox: it's not just geography! *Aust. J. Rural Health* **14**, 95–98.
2. Idriess, I. (1932) *Flynn of the Inland.* Angus & Robertson (Publishers) Pty Ltd., Sydney.
3. National Electronic Decision Support Taskforce. (2002) Report to Health Ministers: Electronic Decision support in Australia. Canberra: National Health Information Management Advisory Council; 2002 November.
4. National Electronic Health Records Taskforce. (2000) *A Health Information Network for Australia*. Commonwealth of Australia, Canberra, 2000 July.
5. Law, J. (1992) Notes on the theory of the actor-network: ordering, strategy, and heterogeneity. *Syst. Prac. Action Res.* **5**, 379–393.
6. Nonaka, I., and Takeuchi, H. (1995) *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press*, Oxford.
7. Latour, B. (1987) *Science In Action. How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge.
8. Tuomi, I. (2000) Internet, innovation, and open source: actors in the network, in *Association of Internet Researchers Conference*, September 15, 2000, Lawrence, Kansas, USA.
9. Olsen, O., and Gtzsche, P. (2006) Screening for breast cancer with mammography (Cochrane Review). The Cochrane Library 2006(2).
10. Rojas, M., Telaro, E., Russo, A., Moschetti, I., Coe, L., Fossati, R., et al. (2000) Follow-up strategies for women treated for early breast cancer (Cochrane Review). The Cochrane Library 2000(2).
11. Bessell, T., Silagy, C., Anderson, J., Hiller, J., and Sansom, L. (2002) Prevalence of South Australia's online health seekers. *Aust. NZ J. Public Health* **26**, 170–173.
12. Ilic, D., Bessell, T., Silagy, C., and Green, S. (2003) Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency. *Hum. Reprod.* **18**, 557–561.
13. Bessell, T., McDonald, S., Silagy, C., Anderson, J., Hiller, J., and Sansom, L. (2002) Do Internet interventions for consumers cause more harm than good? A systematic review. *Health Expect.* **5**: 28–37.
14. Kriege, M., Brekelmans, C., Obdeijn, I., Boetes, C., Zonderland, H., Muller, S., et al. (2006) Factors affecting sensitivity and specificity of screening mammography and MRI in women with an inherited risk for breast cancer. *Breast Cancer Res. Treat.* **100**: 109–119.
15. Williams-Jones, B., and Graham, J. (2003) Actor-Network Theory: a tool to support ethical analysis of commercial genetic testing. *N. Genet. Soc.* **22**: 271–296.
16. Evans, W., and Relling, M. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491.

17. McLeod, H. L., and Evans W. E. (2001) Pharmacogenomics: unlocking the human genome for better drug therapy. *Ann. Rev. Pharm. Toxicol.* **41**, 101–121.
18. Evans, W. E., and Johnson, J. A. (2001) Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Ann. Rev. Genom. Hum. Genet.* **2**, 9–39.
19. Baselga, J., Norton, L., Albanell, J., Kim, Y., and Mendelsohn, J. (1998) Recombinant humanized anti-HER2 antibody (Herceptin) enhances the antitumor activity of paclitaxel and doxorubicin against HER2/neu overexpressing human breast cancer xenografts, *Cancer Res*. **58**, 2825–2831. [published erratum appears in *Cancer Res*. (1999 Apr 15) **59**(8), 2020].
20. Coiera, E., and Westbrook, J. (2006) Should clinical software be regulated? New Australian evaluation guidelines will help inform the debate. *Med. J. Aust.* **184**, 600–601.
21. Slawson, D., Shaughnessy, A., and Bennett, J. (1994) Becoming a medical information master: feeling good about not knowing everything. *J. Fam. Pract*. **38**, 505–513.
22. Eysenbach, G, and Norman, C. (2004) Introduction to CATCH-IT reports: critically appraised topics in communication, health informatics, and technology. *J. Med. Internet Res*. **6**, e49.
23. Shannon, S. (2001) Critically appraised topics (CATs). *Can. Assoc. Radiol. J*. **52**, 286–287.
24. Ebell, M., Barry, H., Slawson, D., and Shaughnessy, A. (1999) Finding POEMs in the medical literature. *J. Fam. Pract* . **48**, 350–355.
25. Guyatt, G., Meade, M., Jaeschke, R., Cook, D., and Haynes, R. (2000) Practitioners of evidence based care. *Br. Med. J*. **320**, 954–955.
26. Delamothe, T. (2000) Quality of websites: kitemarking the west wind. Rating the quality of medical websites may be impossible. *Br. Med. J*. **321**, 843–844.
27. Boyer, C., Selby, M., Scherrer, J., and Appel, R. (1998) The health on the net code of conduct for medical and health websites. *Comput. Biol. Med*. **28**, 603–610.
28. Nater, T., Boyer, C., and Eysenbach, G. (2000) Debate about evaluation and monitoring of sites carrying the HON-Logo. *J. Med. Internet Res*. **2**, e13.
29. Fallis, D., and Fricke, M. (2002) Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home. *J. Am. Med. Inform. Assoc*. **9**, 73–79.
30. Carroll, A., Saluja, S., and Tarczy-Hornoch, P. (2002) Consumer health information on the Internet. *J. Am. Med. Inform. Assoc*. **9**, 402–403.
31. Croft, D., and Peterson, M. (2002) An evaluation of the quality and contents of asthma education on the World Wide Web. *Chest* **121**, 1301–1307.
32. Chen, S. C. W., Blouin, R. A., Mao, Z., Humphries, L. L., Meek, Q. C., Neill J. R., et al. (1996) The cytochrome P450 2D6 (CYP2D6) enzyme polymorphism: screening costs and influence on clinical outcomes in psychiatry. *Clin. Pharmacol. Ther*. **60**, 522–534.

Index