Translational Bioinformatics Series: 1

Series Editor: Xiangdong Wang, MD, PhD, Prof

Yin Yao Shugart *Editor*

Applied Computational Genomics

Springer

Translational Bioinformatics

Series Editor

Xiangdong Wang, MD, PhD
Professor of Clinical Bioinformatics, Lund University, Sweden
Professor of Medicine, Fudan University, China

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Translational Bioinformatics

Series Editor: Xiangdong Wang, MD, PhD, Professor of Clinical Bioinformatics,
Lund University, Sweden;
Professor of Medicine, Fudan University, China

Recently Published and Forthcoming Volumes

Applied Computational Genomics
Editor: Yin Yao Shugart
Volume 1

Pediatric Biomedical Informatics
Editor: John Hutton
Volume 2

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

For further volumes:
http://www.springer.com/series/11057

Translational Bioinformatics

Volume 1

Applied Computational Genomics

Editor: Yin Yao Shugart

Springer

Editor
Yin Yao Shugart
National Institute of Mental Health
Bethesda, MD
USA

ISSN 2213-2775 ISSN 2213-2783 (electronic)
ISBN 978-94-007-5557-4 ISBN 978-94-007-5558-1 (eBook)
DOI 10.1007/978-94-007-5558-1
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012951049

While the advice and information in this book are believed to be true and accurate at the date of publication, 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.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series Foreword

In recognition of the major role informatics plays in accruing, integrating, and analyzing data in the biomedical sciences and translating it into clinical practice, a series of books on Translational Bioinformatics is being created by united scientific forces of the International Society for Translational Medicine (ISTM, www.istmed.org), Journal of Clinical Bioinformatics (JCBi, www.jclinbioinformatics.com), Journal of Clinical and Translational Medicine (CTM, www.clintransmed.com), and Springer Publisher. These will cover topics such as genomics, proteomics, metabolomics, systems immunology, and biomarkers. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research* is an important volume in this series and focuses on core resources in informatics that are necessary to support translational research in a research-intensive children's medical center. One key challenge is implementing interoperable research and clinical IT systems so that data can be exchanged to support translational research.

I, as the Editor of Series Books, am privileged and honored to have Prof. Yin Yao Shugart as the Editor of this special volume. Dr. Shugart has made significant contributions to genomic research in the field of statistical method development and application to human genetics. She has been recognized as a successful researcher and an international leader in the research field of cancer and mental health research.

The part edited by Dr. Shugart strongly focused on the application of these sequencing technologies and places them into the context of techniques such as genome wide association studies (GWAS), family-based linkage analysis, candidate gene based association analysis as well as case-control based association analysis. There are numerous situations where these alternative strategies are closely linked, for example in the case of family-based analyses using data generated on whole exome sequencing (WES) or whole genome sequence (WGS) platforms. The chapters provide plentiful successful examples where researchers have taken advantage of WES or WGS and found casual variants in cancer as well as several Mendelian disorders. Some of these findings have revealed a depth of research investigating disease etiology and the development of promising tools for personalized medicine, enabling earlier diagnosis and targeted treatment. The increasing development of analytical techniques provides a clear route toward greatly improved clinical application of these new sources of discoveries. Machine learning (ML), on the other hand, is an analytical approach which is of increasing importance in this field. In the introductory chapter, the authors discuss the use of ML for disease risk prediction and prognosis and identify successful applications to date.

This book has ten chapters each of which stands alone as a thoughtful mini-review of a specific tool, study design, or broader coverage of a research field. The book is structured in the following manner: Chap. 1 gives an introduction of the nine chapters. The chapters are linked through the cohesive nature of both technological development and statistical knowledge which must work together to improve understanding. Human genetics (or genomics) has experienced many difficulties to reach the stage that the data coming from different platforms can be integrated in an efficient manner and the gained knowledge translated into therapeutic interventions and better prediction tools.

Translational computation genomics, as part of the Springer Series on Translational Bioinformatics, thoroughly discusses the relevant issues to researchers in the field of human genetics, clinical science, and policy making and provides practical guidance to using genomic tools to inform translational research in clinical diagnosis as well as treatment.

Xiangdong Wang M.D., Ph.D.
Professor of Respiratory Medicine
and Director of Biomedical Research Center
Fudan University, Zhongshan Hospital,
Shanghai, China;

Professor (adj) of Clinical Bioinformatics,
Lund University Clinical Science, Lund, Sweden

Contents

Contributors

Andrew Collins Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, University of Southampton, Southampton, UK

Genetic Epidemiology and Genomic Informatics, Faculty of Medicine, University of Southampton, Southampton, UK

Marcella Devoto Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Department of Molecular Medicine, University La Sapienza, Rome, Italy

Tao Feng Case Western Reserve University, Cleveland, OH, USA

Xun Gu Department of Genetics, Development and Cell Biology-Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Timothy J. Jorgensen Department of Radiation Medicine, Georgetown University Medical Center, Washington DC, USA

Francesca Lantieri Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Department of Health Science, Biostatistics Unit, University of Genoa, Genoa, Italy

Mark A. Levenstien Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Shuying Sue Li Statistical Center for HIV/AIDS Research & Prevention (SCHARP), Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Chunyu Liu Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA

Kathleen Ries Merikangas Genetic Epidemiology Research Branch, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Hai-De Qin Unit of Statistical Genomics, Division of Intramural Research Programs, National Institute of Mental Health (NIMH)/NIH, Bethesda, MD, USA

Alan Scott Department of Medicine, Johns Hopkins University, Baltimore, MD, USA

Yin Yao Shugart Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA

Zhixi Su Department of Genetics, Development and Cell Biology-Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Harold Z. Wang Unit of Statistical Genomics, Intramural Research, Program, National Institute of Mental Health, Bethesda, MD, USA

Xinyi Cindy Zhang Biostatistics Program, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Lue Ping Zhao Department of Biostatistics and Department of Epidemiology, School of Public Health and Community Medicine, University of Washington, Seattle, WA, USA

Xiaofeng Zhu Case Western Reserve University, Cleveland, OH, USA

Yangyun Zou Department of Genetics, Development and Cell Biology-Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Abbreviations

ADD/ADHD	Attention Deficit/Hyperactivity Disorder
AIC	Akaike Information Criterion
AML	Acute Myelogenous Leukaemia
ARMD	Age-Related Macular Degeneration
ASD	Autism Spectrum Disorder
BBS	Bardet-Biedl Syndrome
CD	Crohn's Disease
CDs	Related Cognitive Disorders
CGAS	Candidate-Gene Association Studies
CGH	Array Comparative Genome Hybridization
CGP	Cancer Genome Project
CI	Confidence Interval
CNVs	Copy Number Variations
CT	Computerized Tomography
EBV	Epstein Barr Virus
EC2	Elastic Cloud Computing
EE	Estimation Equation
EM	Estimation-Maximization
eQTL	Expression Quantitative Trait Loci
FISH	Fluorescence In Situ Hybridization
GATK	The Genome Analysis Tool Kit
GWAS	Genome-Wide Association Study
HCL	Hairy-Cell Leukemia
HLA	Human Leukocyte Antigen
HWE	Hardy-Weinberg Equilibrium
IBD	Alleles Identical By Descent
IBD	Inflammatory Bowel Disease
ICGC	International Cancer Genome Consortium
ID	Intellectual Disability
IHC	Immunohistochemistry
LCA	Leber Congenital Amaurosis

LCLs	Lymphoblastoid Cell Lines
LD	Linkage Disequilibrium
LMM-EH-PS	Microsoft Linear Mixed Models
LOH	Loss Of Heterozygosity
MAF	Minor-Allele-Frequency
MC	Metachondromatosis
MCMC	Markov Chain Monte Carlo
MHC	Major Histocompatibility Complex
ML	Maximum Likelihood
ML	Machine Learning
MLPA	Multiplex Ligation-Dependent Probe Amplification
mQTL	Metabolic/Metabolite Quantitative Trait Locus
MRI	Magnetic Resonance Imaging
MuTHER	Multiple Tissue Human Expression Resource
NGS	Next-Generation Sequencing
NPC	Nasopharyngeal Carcinoma
OCD	Obsessive-Compulsive Disorder
OR	Odds Ratio
ORWSS	Ration Weighted Sum Statistic
PCR	Polymerase Chain Reaction
PET	Positron Emission Tomography
PGC	Psychiatric Genomics Consortium
pQTL	Protein Qtl
QTL	Quantitative Trait Loci
QTLs	Quantitative Trait Loci
RFLP	Restricted Length Polymorphisms
ROS	Reactive Oxygen Species
RP	Retinitis Pigmentosa
SAM	Sequence Alignment/Map
SD	Segmental Duplications
SED	Spondyloepiphyseal Dysplasia
SGS	Shared Genomic Segment
SKAT	Sequence Kernel Association Test
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeats
SVA	Surrogate Variable Analysis
SVM	Support Vector Machines
SVM	Support Vector Machines
SVs	Structural Variations
TCGA	Cancer Genome Atlas
WES	Whole Exome Sequencing
WSM	Weighted Sum Association Method

Chapter 1
Introduction

Yin Yao Shugart and Andrew Collins

Abstract This chapter presents an overview of the current genomic field, introduces the history of using machine learning for predicative disease studies and provides highlights for all nine chapters which have been collected in this book. The authors also list the critical concepts illustrated by the authors and point out logical connections between different chapters.

Keywords Machine learning • GWAS • Next generation sequencing • Rare variants • eQTLs • Structural mutation • Personalized medicine

1.1 Overview

This is an exciting time for human geneticists focusing on the mechanisms underlying complex traits including cancer, mental health disorders, cardiovascular diseases, diabetes and immune disorders. The current excitement stems from three main technological and analytical developments: (1) the advent of next generation sequencing (NGS) techniques, including whole exome sequencing (WES) and whole genome sequencing (WGS); (2) the development of bioinformatics tools which improve the efficiency and infrastructure for data management and (3) the development of more

Y.Y. Shugart (✉)
Unit of Statistical Genetics, Division of Intramural Research Program, National Institute of Mental Health, 35 Convent Drive, Bethesda, MD 20852, USA
e-mail: kay1yao@mail.nih.gov

A. Collins
Genetic Epidemiology and Genomic Informatics, Faculty of Medicine,
University of Southampton, Duthie Building (808), Southampton General Hospital,
Southampton, SO16 6YD, UK

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_1, © Springer Science+Business Media Dordrecht 2012

powerful statistical tools to analyse large and complex data sets. Despite the technical and conceptual challenges involved in integrating these advances, researchers have already applied NGS approaches to identify disease causal genetic variants and demonstrated functional roles via experimental efforts involving careful validations across various research groups, within ethnically diverse samples and, at times, through animal models. Most importantly, these new developments provide many new opportunities for both experienced and new investigators with fresh knowledge who can develop novel approaches and conceptual models to incorporate progress in molecular genetics, bioinformatics and next generation phenotyping.

This book is focused on the application of these sequencing technologies and places them into the context of techniques such as genome-wide association studies (GWAS), family-based linkage analysis, candidate-gene-based approaches and case–control-based association analysis. There are numerous situations where these alternative strategies are closely linked, for example, in the case of family-based analyses using data generated on WES or WGS platforms. There are already plentiful successful examples where researchers have taken advantage of WES or WGS and found casual variants in cancer as well as several Mendelian disorders. Some of these findings have underpinned a depth of research probing disease aetiology and, more excitingly, the development of novel tools for personalized medicine, enabling earlier diagnosis and targeted cancer treatment. The increasing application and development of sophisticated analytical techniques provides a clear route towards greatly improved clinical application of these new sources of data. Machine learning (ML) is an analytical approach which is of increasing importance in this field. In this introductory chapter, we would like to highlight the use of ML for disease risk prediction and prognosis and identify the scope of successful applications to date. Despite the enthusiasm we feel that evaluation of ML methods in real data sets has been limited so far. We also feel that machine learning approaches can serve as methods of choice for the integration of the ever more complex data sets being generated in this, the era of NGS. It is widely recognized that, in the next few decades, data integration will play an increasingly important role in understanding genome-environment interactions involved in the development of human disorders and the way measured factors modify the function and expression of genes in the genome.

1.2 Machine Learning Approaches: Data Integration for Disease Prediction and Prognosis

Enormous volumes of genomic data encompassing diverse data types (including gene expression, genetic polymorphism, structural mutations, DNA methylation, eQTLs and proteomic data) can now be collected relatively cost-effectively for a large number of patient samples. For inherited disease research, data integration is focused on improving power and accuracy to underpin new discoveries. Integration strategies include meta-analysis where evidence from independent, but essentially similarly structured (homogeneous) data sets is combined across studies.

Meta-analysis has been employed successfully in the context of GWAS (Zeggini et al. 2008) with resulting increased power and consequent novel discoveries.

In a more clinical setting the integration of genomic, proteomic and phenotypic data becomes increasingly important as a route to facilitate diagnosis, enhance treatment and establish prognosis. ML methods are particularly powerful for integrating heterogeneous data sets in both research and clinical settings. ML is an 'artificial intelligence' approach involving a range of statistical and optimization approaches in which computers 'learn' from 'training' data sets to enable predictions about outcomes in further samples. Applications within a clinical setting include numerous examples which have a focus on defining and refining disease diagnosis. In the context of cancer, ML tools have been developed to identify, classify, detect or distinguish tumours (Cruz and Wishart 2006). However, developing applications for ML include disease prediction and prognosis (prediction of disease risk, disease recurrence and survivability) which forms part of the translational research emphasis towards personalized medicine. This field is, however, still in relative infancy and extensive bioinformatic development, validation and demonstrably robust application is required to achieve translational impact. Haskin Fernald et al. (2011) defined the analytical bioinformatics challenges faced in the field of personalized medicine as four main areas: processing voluminous, robust, genome data; interpretation of functional impacts of genome variation; integration of data to establish gene and phenotype relationships in their full complexity; and translation of discoveries into medical practice. ML methods have the potential to become the tools of choice for addressing these challenges as they are demonstrably powerful for integrating voluminous data, refining tools for predicting functional impacts, modelling genotype and phenotype relationships and for integrating genomic and clinical data in a translational manner.

ML methods are particularly useful for large, often noisy and heterogeneous data sets. A range of alternative approaches include multifactor-dimensionality reduction (MDR, Ritchie et al. 2001), neural networks (Motsinger et al. 2008), random forest (Bureau et al. 2005) and support vector machines (SVM, Cortes and Vapnik 1995). Alternative methods have a variety of strengths and limitations which are often application-specific (Upstll-Goddard et al. 2012). Within heterogeneous and complex data sets, ML enables inferences that cannot otherwise be established using conventional statistics which require variable independence and typically include multivariate models based on linear combinations of variables. However, although they are often invaluable in the context of non-linear systems where there is a degree of variable inter-dependence, ML methods are subject to important limitations and careful modelling and evaluation is required to avoid drawing incorrect inferences. A critical limitation is the relationship between the number of variables (features) measured and the number of samples tested. A sample to feature ratio of at least 5–10:1 (Somorjai et al. 2003) is recommended for a robust model. The problem is typified as the 'curse of dimensionality'; the number of features characterizing the data is 'too large' and 'the curse of dataset sparsity'; the number of samples on which these features are measured is 'too small' (Somorjai and Nikulin 1993). Somorjai et al. noted that even when the sample to feature ratio is increased to the recommended level, sparsity of the dataset can still generate misleading results. Similarly, training data sets need to be based on a sufficiently large and representative sample of the whole data set to avoid 'overtraining'.

Support vector machines (SVMs) are state-of-the-art ML methods used for 'supervised learning' to establish training vectors to subsequently classify test samples. Depending on the number of features tested (two or more), the SVM classifier identifies the line, plane or hyperplane that maximally separates two clusters (the 'maximum margin'). The distance between the hyperplane and the closest data points on each side (support vector) is maximized. For example, the genotypes at two or more single nucleotide polymorphisms could be used in a classifier related to good and poor patient survival. Non-linear classifications are achieved using a 'kernel' (which may be a linear, polynomial, sigmoid or radial basis function) which transforms the data into a high-dimensional space. Such kernels can dramatically improve the success of a classifier. For data points that are not readily separated in the model, there is a parameter which reflects the trade-off between minimizing misclassification and maximizing the margin.

SVMs are seeing increasing application in disease prediction and prognosis models. Some recent applications, focusing on refining clinical counselling and treatment pathways, integrate epidemiological data and biomarker expression profiles. For example, Yu et al. (2010) develop a classifier for diabetes based on 14 clinical epidemiological risk measures to predict cases of diabetes and pre-diabetes in a US population. Wan et al. (2012) tested 97 cases with nasopharyngeal carcinoma (NPC) against tissue molecular biomarkers from specific signalling pathways and designed SVM models to refine prognosis measures with 5-year follow-up. The authors established high power for classifying prognosis with potential to direct future therapy. Wang et al. (2012) developed survival classifiers for NPC cases based on expression profiles of 18 tumour-associated biomarkers. The powerful classifier is focused on facilitating counselling and individualized patient management. Schulte et al. (2010) used SVM to predict survival for neuroblastoma based on expression profiles of 430 miRNAs and found highly accurate and independently validated survival prediction. Among the studies that have employed ML with genetic variants as predictors, Listgarten et al. (2004) developed SVM modelling using three SNPs to discriminate breast cancer cases from controls (with 69% predictive power). Jiao et al. (2012) employed ML methods to predict severity of autism spectrum disorder (ASD) based on 29 SNPs from 9 ASD related genes.

The low penetrance and small effect sizes of most 'common' disease variants identified through GWAS currently limit the applicability of this information for disease prediction and prognosis (Moore et al. 2010). To date, hundreds of susceptibility loci for more than 70 diseases have been reported by GWAS. Most variants have modest relative risks, in the range 1.1–1.2, making them very poor disease classifiers and questioning their utility in personalized medicine (Moore and Williams 2009). However, Moore et al. (2010) had argued that GWAS analyses have ignored the full complexity of disease pathobiology, and the linear modelling framework employed considers individual SNPs in isolation from their genomic and environmental context. A more holistic approach recognizes genotype-phenotype relationships in their full complexity and encompasses genetic heterogeneity, gene-gene and gene-environment interactions. These complex interactions are likely to comprise much of the underlying genetic architecture. ML methods have the capability to model this complexity but remain poorly optimized in this context. A particular issue is the development of practical routes for feature selection since it is neither feasible nor desirable to test millions of genomic variants and their higher-order interactions. Moore et al. describe ‘filter’ and ‘wrapper’ strategies for addressing this problem in the context of GWAS data. The hugely voluminous data sets now being established by next generation sequencing make the further development of optimal ML analysis strategies even more pressing if this information is to have translational impact (Szymczak et al. 2009).

1.3 Overview of Chapter Contents

This book has ten chapters each of which stands alone as a thoughtful mini-review of a specific tool, study design or broader coverage of a research field. The book is structured in the following way: Chap. 1 serves as an introduction to the current status of the genomic field and provides highlights of the nine chapters. The chapters are linked through the cohesive nature of both technological development and statistical knowledge which must work together to progress understanding. Human genetics (or genomics) has experienced many difficulties to approach the point where the information generated by different platforms can be integrated in an appropriate manner and the learned knowledge translated into therapeutic interventions or enhanced prediction tools. However, translational computational biology as a field is still young. The need for appropriate integration of data from various platforms including GWAS, WES, WGS and gene expression arises in a wide spectrum of clinical applications. We hope this book provides a flavour of the relevant issues to researchers in the field of human genetics, clinical science, and policymaking and attracts graduate students who are interested in translational research and are willing to contribute to this promising field.

Below, we provide an overview of individual chapters. In Chap. 2, Dr. Merikangas gives a complete review of most important concepts in genetic epidemiology. In this chapter, three co-authors move beyond the traditional risk factors defined by epidemiologists and review breakthroughs in genomics in recent years. Dr. Merikangas gives a precise definition for complex traits and a thorough introduction to genetic epidemiology as a tool for pinpointing the role of genetic factors, as well as environmental factors. The definitions of family studies, twin studies, adoption studies, migration are also reviewed, and issues relevant to the various designs are considered. The chapter illustrates the need for a unified framework for studies of both genetic and environment factors, using narcolepsy as an example. Dr. Merikangas’s work provides a strong foundation for the remainder of the book.

In Chap. 3, Dr. Devoto and her co-authors shared their extensive knowledge with human pedigree data and case–control data. The chapter summarizes important ways in which linkage study designs are applicable in the era of sequence analysis. For instance, the authors describe examples reported by Yokoyama et al. (2011). The investigators’ goal was to identify germ-line mutations predisposing to melanoma using WGS of a single individual in a family with eight affected relatives in three generations as a starting point. They detected a variant in *MITF*, a gene known for being somatically amplified or mutated in some of the individuals who are diagnosed with melanomas. That particular mutation was present in three out of seven affected individuals who were tested in the proband's family. Subsequent testing of additional patients from families of probands with multiple melanomas confirmed the presence of the *MITF* E318K variant in 31 unrelated cases with at least one first- or second-degree relative diagnosed with the same clinical diagnosis. They also conducted linkage analysis of melanoma with E318K under a dominant model with reduced penetrance and a 5% phenocopy rate, and obtained a LOD score of 2.7. Further, the authors confirmed the role of E318K in melanoma using case–control association studies and expression profiling analysis. This successful story indicated that traditional linkage analysis can be usefully employed in the analysis of variants generated by a WGS approach. While these cited examples are highly encouraging, they also remind us of the issue of genetic heterogeneity which is expected to occur frequently in families and complicate the hunt for rare mutations. As editors, we would like to comment that this issue can potentially be addressed by a variety of tactical strategies. One possibility is to treat both genetic penetrance (under a dominant model, for genotype AA and Aa) and phenocopy rate (for genotype aa) as nuisance parameters using a parametric linkage approach. We anticipate that such strategies can work for certain genetic disorders with an underlying unique segregation pattern in a limited number of families. We also foresee an opportunity for new development of statistical methods to identify rare variants in pedigree settings for both qualitative and quantitative traits. (Guo and Shugart 2012).

Distinct from the previous two chapters, Chap. 4 examines research progress in a specific rare disorder. This disease is nasopharyngeal carcinoma (NPC). Dr. Jorgensen et al. conducted a thorough review of all candidate genes related to NPC (note, this was limited to work published in English) and also commented on the findings provided by two GWAS efforts, one by a research group in Taiwan and the other by a group located in Guangzhou, China. Very interestingly, the overlap of genetic markers across all studies was extremely limited, making a meta-analysis of most NPC data sets effectively impossible. However, two GWAS reports gave similar results in terms of the location of the 'significantly associated' variants despite the striking differences in sample sizes in the two different studies (less than 300 cases and controls in Taiwan and approximately 1,500 cases and 1,500 controls in Guangzhou). This observation supports the rationale of conducting GWAS in two high-risk areas for NPC even though the population structure for Taiwanese and Cantonese is quite different. We predict that meta-analysis conducted using these two data sets may reveal novel association signals. Conclusions can be drawn but questions remain. One obvious conclusion is that the HLA region is important. But the question remains: which haplotype(s) are specifically related to the risk of developing NPC? Given our thoughts on the ethnicity difference, we can also speculate that there might be two different haplotypes which 'cause' the NPC phenotype in each population. We envisage that WGS may provide an answer to this question and are eager to see more studies carried out that probe the joint effects of gene and environment involved in development of NPC.

In Chap. 5, Dr. Liu provides an introduction to eQTL studies and thoroughly discusses the implication of successful eQTL mapping. It is known that gene expression levels vary among individuals and can be analysed like other quantitative phenotypes such as height and body mass index. The author summarizes a number of interesting findings from eQTL analysis on human post-mortem brains based on publications which appeared between 2007 and 2012 and concludes that although eQTL mapping in human brain is at its early stage, as a tool, QTL has the potential to identify important disease intermediate phenotypes as well as a route to further understanding of complex diseases. Furthermore, the author describes several commonly used experimental platforms and analytical procedures related to eQTL studies. He also reviews all literatures on mQTL in which DNA methylation levels at specific CpG sites are considered as quantitative traits. More importantly, the author provides a list of databases for QTL mapping results which were built by hard-working scientists who collectively have collated basic scientific knowledge enabling the advancement of personal medicine.

Chapter 6 focus on role of genetic haplotypes in gene prediction. The authors start with a thorough discussion of the theoretical background of haplotype analysis and include their own work on the development of a novel likelihood-based method that builds predictive models using genotypes collected from unrelated individuals and is specifically focused on the HLA region. They propose to firstly construct haplotypes of SNPs with HLA alleles and then build the predictive models based on the constructed haplotypes. They test this new method on the British 1985 birth cohort and show that their prediction accuracy is 10% higher than the method proposed by another group of investigators (Leslie et al. 2008). They further conclude that combining all SNPs and HLA data observed from multi-ethnic populations to build prediction models will lead to further improvement of prediction accuracy. The method proposed in this chapter can be applied to the two existing data sets discussed in Chap. 3. The Taiwanese group reported that an effort was made to build haplotypes using the SNP data. However, the analytical approach used by the investigators was rather traditional. That data set will ultimately be revisited using the method proposed in Chap. 6. It is noteworthy that many immune disorders have been reported to be associated with HLA alleles. The novel method discussed in this chapter and its potential extension is likely to be frequently used in the future.

Chapters 7 and 8 both focus on the topic of WES. Chapter 7 was written by Dr. Collins who has a long track record for conducting association mapping using different genetic markers generated by different platforms. He co-developed the concept of the linkage disequilibrium unit (LDU) with Dr. Morton (Collins and Morton 1998). Dr. Collins expresses the view that exome sequencing in a relatively small number of individuals showing ‘extreme’ phenotypes or more familial subtypes of complex disease may be productive. Dr. Collins also states that WES and WGS both offer the potential to interrogate the cumulative impact of the numerous rare variants presumed to underlie a substantial proportion of complex disease susceptibility. On the other hand, the author comments that both WES and WGS will yield enormous amounts of data and pose many analytical challenges. While the cutting edge sequencing technologies provide high-resolution measurements of biological quantities, these new biotechnologies also raise novel statistical and computational challenges in areas such as image analysis, base-calling and read-mapping in initial analysis together with peak-finding. Furthermore, the author also introduces the main statistical methods that can be used to analyse both rare variants and CNVs. Readers who are eager to grasp analytical concepts relating to *de novo* variants**,** the behaviours of rare variants in families versus large cohorts and technical details related to sequencing alignment and variant calling, as well as data management, will find this chapter useful. The in-depth statistical framework for rare variants analysis can be found in Chap. 8 in which all currently analytical strategies on rare variant hunting are discussed. Drs. Feng and Zhu first explain the idea of 'collapsing' and then provide mathematical algorithms on all methods including weighted sum association method (WSM) (Feng et al. 2011), pooled association tests for rare variants, data-adaptive aSUM test, alpha-test (Han and Pan 2010), sequence kernel association test (SKAT) (Wu et al. 2011), and odds ration weighted sum statistic (ORWSS) (Price et al. 2010). Furthermore, they give a description for a general framework developed by Lin and Tang (2011) for the purpose of detecting disease associations with rare variants in sequencing studies.

Chapter 9 covers a very important area of human genetics, which is copy number variation. Dr. Gu et al. discuss five ambitious topics which are CNV and segmental duplications, rate of segmental duplication and rate of CNV essentiality (or dispensability) in duplicate genes, transcriptional regulation divergence following gene duplication, and gene duplication and environmental adaptation. The authors use abundant experimental material generated from their own lab. They introduce us to the concepts using many vivid examples, revealing a combination of exciting and intriguing findings, and help us to interpret the link between copy number variations and complex traits in humans such as psychiatric disorders.

Finally, Chap. 10 aims to serve as a bridge to connect the details provided by previous chapters and the goal of translational research. Translational medicine, based on understanding genomic architecture, is expected to help overcome the diagnostic difficulties for both cancer and mental illnesses by illuminating common patterns of genetic change in specific subtypes of disease. The invention of WGS, WES, RNA-seq, Met-seq and ChIp-seq have resulted in powerful ways to interrogate the genome. NGS is already transforming both genetic research clinical practice by providing much better resolution of the underlying genetic mechanisms for cancer, mental illnesses and a spectrum of other diseases. Diagnoses are already being made with greater precision, and treatments are being individualized based on a more thorough understanding of each patient. In this chapter, Qin and Shugart focus on this rapidly progressing field, covering topics on the application of NGS to translational medicine in oncology and psychiatry. At the end of this chapter, the authors emphasize that our goal as geneticists and molecular biologists is to make certain that we can understand genetic variation at a level where it will make a significant contribution to people's health. Therefore, it is imperative to develop a new system for coordinating basic discovery with patient medical records ways that facilitate smooth transition to personalized medicine in the future. Epidemiologists, statisticians, bioinformaticians, biologists and physicians must work together more closely than ever before to fulfil this ambitious goal.

References

Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol. 2005;28:171–82.

Collins A, Morton NE. Mapping a disease locus by allelic association. Proc Natl Acad Sci USA. 1998;95:1741–45.

Cortes C, Vapnik V. Support vector networks. Mach Learn. 1995;20:273–97.

Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006;2:59–77.

Feng T, Elston RC, Zhu XF. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). Genet Epi. 2011;35:398–409.

Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011;27(13):1741–8.

Guo W, Shugart YY. Detecting rare variants for quantitative traits using nuclear families. Hum Hered. 2012;73:148–58.

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 2010;70:42–54.

Jiao Y, Chen R, Ke X, Cheng L, Chu K, Lu Z, Herskovits EH. Single nucleotide polymorphisms predict symptom severity of autism spectrum disorder. J Autism Dev Disord. 2012;42(6):971–83.

Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011;89:354–67.

Listgarten J, Damaraju S, Poulin B, et al. Predictive models for breast cancer susceptibility from single nucleotide polymorphisms. Clin Cancer Res. 2004;10:2725–37.

Moore JH, Asselbergs FW, William SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010;26(4):445–56.

Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet. 2009;85:309–20.

Motsinger-Reif A, Dudek SM, Hahn LW, et al. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol. 2008;32:325–40.

Price AL, Kryukov GV, de Bakker PI, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86:832–8.

Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69:138–47.

Schulte JH, Schowe B, Mestdagh P, Kaderali L, Kalaghatgi P, Schlierf S, Vermeulen J, Brockmeyer B, Pajtler K, Thor T, de Preter K, Speleman F, Morik K, Eggert A, Vandesompele J, Schramm A. Accurate prediction of neuroblastoma outcome based on miRNA expression profiles. Int J Cancer. 2010;127(10):2374–85.

Somorjai RL, Nikulin A. The curse of small sample sizes in medical diagnosis via MR spectroscopy. In: Proceedings of the Society for Magnetic Resonance in Medicine. Twelfth annual scientific meeting, New York; 1993. pp. 685.

Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics. 2003;19:1484–91.

Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV. Machine learning in genome-wide association studies. Genet Epidemiol. 2009;33:S51–7.

Upstll-Goddard R, Eccles D, Fliege J, Collins A. Machine learning approaches for the discovery of gene-gene interactions in disease data. Brief Bioinf. 2012; doi:10.1093.

Wan XB, Zhao Y, Fan XJ, Cai HM, Zhang Y, Chen MY, Xu J, Wu XY, Li HB, Zeng YX, Hong MH, Liu QT. Molecular prognostic prediction for locally advanced nasopharyngeal carcinoma by support vector machine integrated approach. PLoS One. 2012;7(3):e31989.

Wang HY, Sun BY, Zhu ZH, Chang ET, To KF, Hwang JSG, et al. Eight-signature classifier for prediction of nasopharyngeal carcinoma survival. J Clin Oncol. 2012;29(34):4516–24.

Wu MC, Lee S, Cai T, et al. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). Am J Hum Genet. 2011;89:82–93.

Yokoyama S, Woods SL, Boyle GM, et al. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. Nature. 2011;480:99–103.

Yu W, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10:16.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008;40(5):638–45.

Chapter 2
Concepts of Genetic Epidemiology

Kathleen Ries Merikangas

Abstract The major aim of this chapter is to provide an overview of the field of genetic epidemiology and its relevance to the identification of the causes and risk factors for human diseases. The most important goal of the methods of genetic epidemiology is to elucidate the joint contribution of genes and environmental exposures to the etiology of complex diseases. The key study designs used to achieve this goal including family, twin, adoption, and migration studies are summarized. The field of genetic epidemiology is expected to have increasing importance with advances in molecular genetics.

Keywords Genetics • Epidemiology • Family studies • Twin studies • Adoption studies • Migration studies

2.1 Introduction: Genetic Epidemiology

Genetic epidemiology is defined as the study of the distribution of and risk factors for diseases and genetic and environmental causes of familial resemblance. Genetic epidemiology focuses on how genetic factors and their interactions with other risk factors increase vulnerability to, or protection against, disease (Beaty 1997). Genetic epidemiology employs traditional epidemiologic study designs to explain aggregation in groups as closely related as twins or as loosely related as migrant cohorts. Epidemiology has developed sophisticated designs and analytic methods for identifying

K.R. Merikangas, Ph.D. (✉)
Genetic Epidemiology Research Branch, Intramural Research Program,
National Institute of Mental Health, National Institutes of Health,
35 Convent Drive, MSC#3720, Bethesda, MD 20892, USA
e-mail: Kathleen.merikangas@nih.gov

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_2, © Springer Science+Business Media Dordrecht 2012

disease risk factors. With increasing progress in gene identification, these methods have been extended to include both genetic and environmental factors (MacMahon and Trichopoulos 1996; Kuller 1979). In general, study designs in genetic epidemiology either control for genetic background while letting the environment vary (e.g., migrant studies, half siblings, separated twins) or control for the environment while allowing variance in the genetic background (e.g., siblings, twins, adoptees/nonbiological siblings). Investigations in genetic epidemiology are typically based on a combination of study designs including family, twin, and adoption studies.

2.1.1 Family Studies

Familial aggregation is generally the first source of evidence that genetic factors may play a role in a disorder. The most common indicator of familial aggregation is the relative risk ratio, computed as the rate of a disorder in families of affected persons divided by the corresponding rate in families of controls. The patterns of genetic factors underlying a disorder can be inferred from the extent to which patterns of familial resemblance adhere to the expectations of Mendelian laws of inheritance. The degree of genetic relatedness among relatives is based on the proportion of shared genes between a particular relative and an index family member or proband. First-degree relatives share 50% of their genes in common, second-degree relatives share 25% of their genes in common, and third-degree relatives share 12.5% of their genes in common. If familial resemblance is wholly attributable to genes, there should be a 50% decrease in disease risk with each successive increase in degree of relatedness, from first to second to third and so forth. This information can be used to derive estimates of familial recurrence risk within and across generations as a function of population prevalence (λ) (Risch 1990b). Whereas λ tends to exceed 20 for most autosomal dominant diseases, values of λ derived from family studies of many complex disorders tend to range from 2 to 5. Diseases with strong genetic contributions tend to be characterized by 50% decrease in risk across successive generations. Decrease in risk according to the degree of genetic relatedness can also be examined to detect interactions between several loci. If the risk to second- and third-degree relatives decreases by more than 50%, this implies that more than a single locus must contribute to disease risk and that no single locus can largely predominate.

The major advantage of studying diseases within families is that disease manifestations are more likely to result within families than they are across families from the same underlying etiologic factors. Family studies are therefore more effective than between family designs in examining the validity of diagnostic categories because they more accurately assess the specificity of transmission of symptom patterns and disorders. Data from family studies can also provide evidence regarding etiologic or phenotypic heterogeneity. Phenotypic heterogeneity is suggested by variable expressivity of symptoms of the same underlying risk factors, whereas etiologic heterogeneity is demonstrated by common manifestations of expression of different etiologic factors between families. Moreover, the family study method permits assessment of associations between disorders by evaluating specific patterns of co-segregation of two or more disorders within families (Merikangas 1990).

2.1.2 Twin Studies

Twin studies that compare concordance rates for monozygotic twins (who share the same genotype) with those of dizygotic twins (who share an average of 50% of their genes) provide estimates of the degree to which genetic factors contribute to the etiology of a disease phenotype. A crude estimate of the genetic contribution to risk for a disorder is calculated by doubling the difference between the concordance rates for monozygous and dizygous twin pairs. Modern genetic studies employ path analytic models to estimate the proportion of variance attributable to additive genes, common environment, and unique environment. There are several other applications of the twin study design that may inform our understanding of the roles of genetic and environmental risk factors for disease. First, twin studies provide information on the genetic and environmental sources of sex differences in a disease. Second, environmental exposures may be identified through comparison of discordant monozygotic twins. Third, twin studies can be used to identify the genetic mode of transmission of a disease by inspection of the degree of adherence of the difference in risk between monozygotic and dizygotic twins to the Mendelian ratio of 50%. Fourth, twin studies may contribute to enhancing the validity of a disease through inspection of the components of the phenotypes that are most heritable. The twin family design is one of the most powerful study designs in genetic epidemiology because it yields estimates of heritability but also permits evaluation of multigenerational patterns of expression of genetic and environmental risk factors.

2.1.3 Adoption Studies

Adoption studies have been the major source of evidence regarding the joint contribution of genetic and environmental factors to disease etiology. Adoption studies either compare the similarity between an adoptee and his or her biological versus adoptive relatives or the similarity between biological relatives of affected adoptees with those of unaffected or control adoptees. The latter approach is more powerful because it eliminates the potentially confounding effect of environmental factors. Similar to the familial recurrence risk, the genetic contribution in adoption studies is estimated by comparing the risk of disease to biological versus adoptive relatives or the risk of disease in biological relatives of affected versus control adoptees. These estimates of risk are often adjusted for sex, age, ethnicity, and other factors that may confound the links between adoption status and an index disease.

With the recent trends toward selective adoption and the diminishing frequency of adoptions in the USA, adoption studies are becoming less feasible methods for identifying genetic and environmental sources of disease etiology (National

Adoption Information Clearinghouse 2007). However, the increased rate of reconstituted families (families comprised of both siblings and half siblings) may offer a new way to evaluate the role of genetic factors in the transmission of complex disorders. Genetic models predict that half siblings should have a 50% reduction in disease risk compared to that of full siblings. Deviations from this risk provide evidence for either polygenic transmission, gene-environment interaction, or other complex modes of transmission.

2.1.4 Migration Studies

Migrant studies are perhaps the most powerful study design to identify environmental and cultural risk factors. When used to study Asian immigrants to the USA, this study design demonstrated the significant contribution of the environment to the development of many forms of cancer and heart disease (Kolonel et al. 2004). One of the earliest controlled migrant studies evaluated rates of psychosis among Norwegian immigrants to Minnesota compared to native Minnesotans and native Norwegians (Ödegaard 1932). A higher rate of psychosis was found among the immigrants compared to both the native Minnesotans and Norwegians and was attributed to increased susceptibility to psychosis among the migrants who left Norway. It was found that migration selection bias was the major explanatory factor, rather than environmental exposure in the new culture. The application of migration studies to the identification of environmental factors is only valid if potential bias attributed to selection is considered. Selection bias has been tested through comparisons of factors that may influence a particular disease of interest in a migrant sample and a similar sample that did not migrate.

2.2 Applications of Genetic Epidemiology to Gene Identification

There is a widespread consensus among geneticists and epidemiologists on the importance of epidemiology to the future of genetics and on the conclusion that the best strategy for susceptibility risk factor identification for common and complex disorders will ultimately involve large epidemiologic studies from diverse populations(Peltonen and McKusick 2001; Khoury and Little 2000; Yang and Khoury 1997; Merikangas 2003; Merikangas and Risch 2003; Risch 1990a). It is likely that population-based association studies will assume increasing importance in translating the products of genomics to public health. There are several reasons that population-based studies are critical to current studies seeking to identify genes underlying complex disorders. First, the frequency of newly identified polymorphisms, whether SNPs or other variants such as copy number variations (CNVs), especially in particular population subgroups, is not known. Second, current knowledge of genes as risk factors is based nearly exclusively on clinical and nonsystematic samples. Hence, the significance of the susceptibility alleles that have been identified for cancer, heart disease, diabetes, and other common disorders is unknown in the population at large. In order to provide accurate risk estimates, the next stage of research needs to move beyond samples identified through affected individuals to the population as a whole. Third, identification of risk profiles will require large samples to assess the significance of vulnerability genes with relatively low expected population frequencies. Fourth, similar to the role of epidemiology in quantifying risk associated with traditional disease risk factors, applications of human genome epidemiology can provide information on the specificity, sensitivity, and impact of genetic tests to inform science and individuals (Khoury and Little 2000).

2.2.1 Samples

The shift from systematic large-scale family studies to linkage studies has led to the collection of families according to very specific sampling strategies (e.g., many affected relatives, affected sibling pairs, affected relatives on one side of the family only, and availability of parents for study) in order to maximize the power of detecting genes according to the assumed model of familial transmission. Despite the increase in power for detecting genes, these sampling approaches have diminished the generalizability of the study findings and contribute little else to the knowledge base if genes are not discovered. Future studies will attempt to collect both families and controls from representative samples of the population so that results can be used to estimate population risk parameters and to examine the specificity of endophenotypic transmission and so results can be generalized to whole populations.

2.2.2 Selection of Controls

The most serious problem in the design of association studies is the difficulty of selecting controls that are comparable to the cases on all factors except the disease of interest (Wacholder et al. 2000; Ott 2004). Controls should be drawn from the same population as cases and must have the same probability of exposure (i.e., genes) as cases. Controls should be selected to ensure the validity rather than the representativeness of a study. Failure to equate cases and controls may lead to confounding (i.e., a spurious association due to an unmeasured factor that is associated with both the candidate gene and the disease). In genetic case–control studies, the most likely source of confounding is ethnicity because of differential gene and disease frequencies in different ethnic subgroups. The matching of controls to cases on ethnic background is largely based on self-report; several methods are used to screen for and exclude subjects with substantial differences in ancestry.

2.2.3 Risk Estimation

Because genetic polymorphisms involved in complex diseases are likely to be nondeterministic (i.e., the marker neither predicts disease nor non-disease with certainty), traditional epidemiologic risk factor designs can be used to estimate the impact of these genetic polymorphisms. Increased attention to alleles as a part of risk equations in epidemiology will likely resolve the contradictory findings from studies that have generally employed solely environmental risk factors, such as diet, smoking, and alcohol use. Likewise, the studies that seek solely to identify small risk alleles will continue to be inconsistent because they do not consider the effects of nongenetic biological parameters or environmental factors that contribute to the diseases of interest.

There are several types of risk estimates that are used in public health. The most common is *relative risk*, defined as the magnitude of the association between an exposure and disease. It is independent of the prevalence of the exposure. The *absolute risk* is the overall probability of developing a disease in an individual or in a particular population (Gordis 2000). The *attributable risk* is the difference in the risk of the disease in those exposed to a particular risk factor compared to the background risk of a disease in a population (i.e., in the unexposed). The *population attributable risk* relates to the risk of a disease in a total population (exposed and unexposed) and indicates the amount the disease can be reduced in a population if an exposure is eliminated. The population attributable risk depends on the prevalence of the exposure or, in the case of risk alleles, the allele frequency. Genetic attributable risk would indicate the proportion of a particular disease that would be eliminated if a particular gene or genes were not involved in the disease. For example, the two vulnerability alleles for Alzheimer's disease include the very rare but *deterministic alleles* in the β-amyloid precursor, presenilin-1, and presenilin-2 genes, which signal a very high probability of the development of Alzheimer's disease, particularly at a young age, and the *susceptibility* allele ε4 in the apolipoprotein-E gene (APOE ε4) (Tol et al. 1999). The apolipoprotein-E ε4 (APOE ε4) allele has been shown to increase the risk of Alzheimer's disease in a dose-dependent fashion. Using data from a large multiethnic sample collected by more than 40 research teams, Farrer (Farrer et al. 1997) reported a 2.6–3.2 greater odds of Alzheimer's disease among those with one copy and 14.9 odds of Alzheimer's disease among those with two copies of the APOE ε4 allele. Moreover, there was a significant protective effect among those with the ε2/ε3 genotype. As opposed to the deterministic mutations, the APOE ε4 allele has a very high population attributable risk because of its high frequency in the population. The APOE ε4 allele is likely to interact with environmental risk and protective factors (Kivipelto et al. 2001; Kivipelto et al. 2002). The population risk attributable to these mutations is quite low because of the very low population prevalence of disease associated with these alleles. This model of combination of several rare deterministic alleles in a small subset of families and common alleles with lower relative risk to individuals but high population attributable risk is likely to apply to many other complex diseases as well. Genome-wide association studies have now identified genes for more than 300 diseases and traits, such as coronary artery disease, Crohn's disease, rheumatoid arthritis, and type 1 and type 2 diabetes (Wellcome Trust Case Control Consortium 2007) with 1,291 publications by the end of 2011 (www.genome.gov/gwastudies). Those genetic variants appear to confer only modest increases in disease risk (ORs between 1.2 and 1.5) compared to other established risk factors for common chronic diseases.

2.2.4 Identification of Environmental Factors

In parallel with the identification of susceptibility alleles, it is important to identify environmental factors that operate either specifically or nonspecifically on those with susceptibility to complex disorders in order to develop effective prevention and intervention efforts. Langholz et al. (1999) describe some of the world's prospective cohort studies that may serve as a basis for studies of gene-disease associations or gene-environment interactions. There is increasing evidence that gene-environment interaction will underlie many of the complex human diseases. Some examples include inborn errors of metabolism, individual variation in response to drugs (Nebert 1999), substance use disorders (Heath et al. 2001; Rose et al. 2000), and the protective influence of a deletion in the CCR5 gene on exposure to HIV (Michael 1999). In prospective studies, however, few environmental exposures have been shown to have an etiologic role in complex disorders (Eaton 2004). Over the next decades, it will be important to identify and evaluate the effects of specific environmental factors on disease outcomes and to refine measurement of environmental exposures to evaluate the specificity of effects. Study designs and statistical methods should focus increasingly on the nature of the relationships between genetic and environmental factors, particularly epistasis and gene-environment interaction (Yang and Khoury 1997; Ottman 1990; Beaty and Khoury 2000). For example, recent breakthroughs in identifying the mechanisms for hypocretin deficiency as the causal mechanism in narcolepsy occurred through a convergence of epidemiologic studies that documented a recent surge in incidence among those exposed to H1N1 virus or vaccine, successful application of genome-wide association studies that implicated specific autoimmune mechanisms (i.e., the T-cell receptor α polymorphism), and specificity of the findings for the phenotype of narcolepsy with cataplexy rather than narcolepsy alone (Kornum et al. 2011).

2.3 Applications, Impact, and Future Directions

The advances in bioinformatics and statistical methods described in the following chapters will be critical to translation of progress in molecular genetics to human diseases. Genetic epidemiologic approaches, particularly the family study design, will have renewed importance in facilitating integration between methodological developments and human diseases. Despite the long history of information provided by family studies regarding the genetic architecture of Mendelian diseases as well as heterogeneity of complex diseases such as breast cancer (Claus et al. 1993) and diabetes (Hawa et al. 2002), the family study approach has largely been abandoned in psychiatry in favor of very large case–control studies of diagnosed patients from clinical samples or registries. Yet, family studies still have an essential role in identifying cross-generational transmission of phenotypes and genotypes. Family-based studies will be even more valuable with application of advances in molecular biology to inform interpretation of sequencing data and to distinguish *de novo* from heritable structural variants. Based on increasing awareness of the neglect of family studies for risk prediction, even in the absence of specification of disease genetic architecture, the US surgeon general has launched a national public health campaign to encourage all American families to learn more about their family health history (http://www.hhs.gov/familyhistory/). A positive family history remains a more potent predictor of disease vulnerability than nearly all other risk factors combined (Meigs et al. 2008). Moreover, since genetic factors, common environmental exposure, and sociocultural factors have been shown to jointly contribute to disease etiology, family history may ultimately have greater explanatory power than genes in predicting risk, particularly if genetic influences are weak.

Progress in genomics has far outstripped advances in our understanding of many of the complex multifactorial human disorders and their etiologies. Technical advances and availability of rapidly expanding genetic databases provide extraordinary opportunities for understanding disease pathogenesis. Over the next decade, increasing understanding of the complex mechanisms through which genetic risk factors influence disease should enhance the clinical utility of genetics. The above issues regarding sampling, complexity of the links between genes and environmental factors in multifactorially determined complex diseases, and phenotypic heterogeneity also highlight the complexity of etiology of complex human diseases. This work demonstrates that predictions that human genomics would lead to a radical transformation of medical practice were overly optimistic. In fact, Varmus (Varmus 2002) concluded that despite the journalistic hyperbole, the sequencing of the human genome is unlikely to lead either to a radical transformation of medical practice or even to an information-based science that can predict with certainty future diseases and effective treatment interventions. Therefore, despite the extraordinary opportunity for understanding disease pathogenesis afforded by the technical advances and availability of rapidly expanding genetic databases, it is unlikely that we will soon experience the light speed progress of genomics in understanding, treating, or preventing many of the multifactorial complex human diseases.

The chasm between genetic information and clinical utility should gradually close as we develop new methods and tools in human genetic and clinical research to maximize the knowledge afforded by the exciting advances in genomics. Increased integration of advances in basic sciences and genomics along with information from population-based studies and longitudinal cohorts; innovations in our conceptualizations of the disease etiology, particularly the role of infectious agents; and the identification of specific risk and protective factors will lead to more informed intervention strategies. As we learn more about the role of genes as risk factors, rather than as the chief causes of common human diseases, it will be essential to provide accurate risk estimation and to inform the public of the need for population-based integrated data on genetic, biological, and environmental risk factors.

The goal of genomics research is ultimately prevention, the cornerstone of public health. An understanding of the significance of genetic risk factors and proper interpretations of their meaning for patients and their families will ultimately become part of clinical practice. Clinicians will become increasingly involved in helping patients to comprehend the meaning and potential impact of genetic risk for complex disorders. As our knowledge of the role of genetic risk factors in advances, it will be incumbent upon clinicians to become familiar with knowledge gleaned from genetic epidemiologic and genomics research. In the meanwhile, use of recurrence risk estimates from family studies best predicts the risk of the development of complex disorders.

References

Beaty TH. Evolving methods in genetic epidemiology. I. Analysis of genetic and environmental factors in family studies. Epidemiol Rev. 1997;19(1):14–23.

Beaty TH, Khoury MJ. Interface of genetics and epidemiology. Epidemiol Rev. 2000;22(1):120–5.

Claus EB, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. Breast Cancer Res Treat. 1993;28(2):115–20.

Eaton WW. Risk factors for mental health disorders (Unpublished Report). National Institute of Mental Health; 2004.

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer disease meta analysis consortium. JAMA. 1997;278(16):1349–56.

Gordis L, editor. Epidemiology. 2nd ed. Philadelphia: WB Saunders; 2000.

Hawa MI, Beyan H, Buckley LR, Leslie RD. Impact of genetic and non-genetic factors in type 1 diabetes. Am J Med Genet. 2002;115(1):8–17. doi:10.1002/ajmg.10339.

Heath AC, Whitfield JB, Madden PA, Bucholz KK, Dinwiddie SH, Slutske WS, Bierut LJ, Statham DB, Martin NG. Towards a molecular epidemiology of alcohol dependence: analysing the interplay of genetic and environmental risk factors. Br J Psychiatry Suppl. 2001;40:s33–40.

Khoury MJ, Little J. Human genome epidemiologic reviews: the beginning of something HuGE. Am J Epidemiol. 2000;151(1):2–3.

Kivipelto M, Helkala EL, Hanninen T, Laakso MP, Hallikainen M, Alhainen K, Soininen H, Tuomilehto J, Nissinen A. Midlife vascular risk factors and late-life mild cognitive impairment: a population-based study. Neurology. 2001;56(12):1683–9.

Kivipelto M, Helkala EL, Laakso MP, Hanninen T, Hallikainen M, Alhainen K, Iivonen S, Mannermaa A, Tuomilehto J, Nissinen A, Soininen H. Apolipoprotein E epsilon4 allele, elevated midlife total cholesterol level, and high midlife systolic blood pressure are independent risk factors for late-life Alzheimer disease. Ann Intern Med. 2002;137(3):149–55.

Kolonel LN, Altshuler D, Henderson BE. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. Nat Rev Cancer. 2004;4(7):519–27. doi:10.1038/nrc1389.

Kornum BR, Faraco J, Mignot E. Narcolepsy with hypocretin/orexin deficiency, infections and autoimmunity of the brain. Curr Opin Neurobiol. 2011;21(6):897–903. doi:S0959-4388(11)00147-4.

Kuller LH. The role of population genetics in the study of the epidemiology of cardiovascular risk factors. Prog Clin Biol Res. 1979;32:489–95.

Langholz B, Rothman N, Wacholder S, Thomas DC. Cohort studies for characterizing measured genes. J Natl Cancer Inst Monogr. 1999;26:39–42.

MacMahon B, Trichopoulos D, editors. Epidemiology: principles and methods. Boston: Little Brown and Company; 1996.

Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino Sr RB, Cupples LA. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008;359(21):2208–19. doi:359/21/2208.

Merikangas KR, editor. Comorbidity of mood and anxiety disorders. Washington, DC: American Psychiatric Press Inc; 1990.

Merikangas KR, editor. Genetic epidemiology of substance-use disorders, in biological psychiatry. Chichester: Wiley; 2003.

Merikangas KR, Risch N. Genomic priorities and public health. Science. 2003;302(5645):599–601. doi:10.1126/science.1091468.

Michael NL. Host genetic influences on HIV-1 pathogenesis. Curr Opin Immunol. 1999;11(4):466–74. doi:10.1016/s0952-7915(99)80078-8.

National Adoption Information Clearinghouse. The adoption home study process, 2007, available online at http://naic.acf.hhs.gov/pubs/f_homstu.cfm

Nebert DW. Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? Clin Genet. 1999;56(4):247–58.

Ödegaard Ö, editor. Emigration and insanity: a study of mental disease among the Norwegian born population of Minnesota. Copenhagen: Levin & Munksgaards; 1932.

Ott J. Association of genetic loci: replication or not, that is the question. Neurology. 2004;63(6):955–8.

Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol. 1990;7(3):177–85. doi:10.1002/gepi.1370070302.

Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. Science. 2001;291(5507):1224–9.

Risch N. Genetic linkage and complex diseases, with special reference to psychiatric disorders. Genet Epidemiol. 1990a;7(1):3–16; discussion 17–45. doi:10.1002/gepi.1370070103.

Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet. 1990b;46(2):222–8.

Rose RJ, Dick DM, Viken RJ, Kaprio J. Gene-environment interaction in patterns of adolescent drinking: regional residency moderates longitudinal influences of alcohol use. Alcohol Clin Exp Res. 2000;25:637.

Tol J, Roks G, Slooter AJ, van Duijn CM. Genetic and environmental factors in Alzheimer's disease. Rev Neurol (Paris). 1999;155 Suppl 4:S10–16.

Varmus H. Getting ready for gene-based medicine. N Engl J Med. 2002;347(19):1526–7. doi:10.1056/NEJMe020119.

Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst. 2000;92(14):1151–8.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78. doi:10.1038/nature05911.

Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev. 1997;19(1):33–43.

Chapter 3
Integration of Linkage Analysis and Next-Generation Sequencing Data

Francesca Lantieri, Mark A. Levenstien, and Marcella Devoto

Abstract Genetic mapping by linkage analysis has been for many years the first step in the identification of genes responsible for rare Mendelian disorders. When the focus of genetic research shifted toward the study of the more complex common disorders, alternative approaches such as association studies were shown to be more successful in identifying common variants of small effect that are in part responsible for susceptibility to such conditions. Recent advances in technologies that make feasible the sequencing of whole exomes or genomes have renewed interest in the identification of rare variants, which are in principle amenable to being detected by linkage analysis. As a result, linkage analysis and family-based studies in general are being reexamined as an aid to filter and validate results of whole exome and whole genome sequencing experiments. This chapter will describe a few

F. Lantieri
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

Department of Health Science, Biostatistics Unit, University of Genoa, Genoa, Italy

M.A. Levenstien
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

M. Devoto (✉)
Division of Human Genetics, The Children's Hospital of Philadelphia,
3615 Civic Center Blvd., ARC 1002, Philadelphia, PA, 19104, USA

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA

Department of Biostatistics and Epidemiology, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA

Department of Molecular Medicine, University La Sapienza,Rome, Italy
e-mail: devoto@chop.edu

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_3, © Springer Science+Business Media Dordrecht 2012

representative papers that have incorporated linkage analysis and its results in the design, execution, and interpretation of whole genome or whole exome sequencing studies.

Keywords Linkage analysis • Rare variants • Family-based studies • Whole exome sequencing • Whole genome sequencing

3.1 Introduction

Linkage analysis and family-based tests have been a workhorse of genetic mapping for Mendelian disease gene identification. From the beginning of the 1980s, the combination of increasingly dense DNA marker maps and powerful software tools implementing such tests have led to the identification of the genes responsible for thousands of Mendelian disorders (Botstein and Risch 2003). When the focus of genetic research shifted from the rare, highly penetrant monogenic diseases to the more common, complex ones, it became evident that linkage analysis was underpowered to detect the common risk variants with small effects expected under the common disease/common variant hypothesis (Risch and Merikangas 1996). Instead, genome-wide association studies (GWAS) in case–control datasets have led to the identification of many such variants in a variety of different disorders and traits (http://www.genome.gov/gwastudies/). At the same time, it has become clear that common variants do not explain all the genetic susceptibility to such traits, and evidence has been accumulating that rare, possibly higher penetrant variants also underlie susceptibility to common complex traits (Manolio et al. 2009). In addition, many Mendelian disorders are too rare for the linkage analysis approach alone to work, and thus the corresponding genes still remain undetected.

The advent of massive parallel sequencing and the ability to sequence the whole exome or even genome of individuals at a relatively low cost has made the discovery of all variants present in an individual or family technically possible. These advances can lead to successful disease gene identification, as demonstrated initially in a few Mendelian disorders (Bamshad et al. 2011) and more recently in some complex ones (Zeggini 2011). However, analysis of whole exome or whole genome sequence (WES or WGS) data poses noticeable bioinformatics and statistical challenges, and the identification of the true risk variants among the many detected by such experiments has often been compared to finding the classic needle in a haystack. Every possible piece of information that can be used to facilitate such effort should be considered and incorporated into the analysis, and in this respect, analysis of the segregation of candidate risk variants in family members of affected individuals has been suggested as particularly useful (Cirulli and Goldstein 2010). In fact, linkage analysis can inform interpretation of WES or WGS data both by indicating regions of the genome with higher a priori chance of including the risk variants when results of linkage studies on the disease of interest are already available and a posteriori by limiting further evaluation of candidate variants detected through sequencing only to those that show co-segregation (i.e., linkage) to the disease or trait in families of affected individuals.

In this chapter, we will describe ways in which linkage analysis and family-based data have been incorporated into WES or WGS experiments that have led to the identification of new disease gene variants. We will also discuss methods that have been proposed for the specific purpose of integrating the use of family-based data in WES or WGS analysis.

3.2 Linkage Analysis in WES/WGS Studies and Identification of Disease Genes

Co-segregation of variants identified by WES/WGS and the disease phenotype in affected relatives is an obvious filter to impose on results of such experiments to reduce the number of candidate variants. As such, numerous studies have used this relatively simple strategy (Ng et al. 2010), which, however, does not take full advantage of the power of linkage analysis to identify candidate regions by modeling the disease mode of inheritance using allele frequency, reduced penetrance, and phenocopy rate. We will not review such studies, as their number is already large and increasing, and the approach relatively straightforward (i.e., remove from further consideration all variants that are not shared by affected relatives). Rather, we will review a few experimental studies that have integrated a formal linkage analysis with their sequencing experiments at various stages. As with all reviews, this list is necessarily limited, but we hope it will still illustrate different ways in which investigators are taking advantage of the power of linkage analysis in their sequencing experiments for the identification of disease genes.

A good proof of principle of the power of both the linkage and the whole exome sequencing approaches is the study of Bowne et al. (2011). These authors investigated an Irish family with autosomal dominant retinitis pigmentosa (adRP) by linkage analysis on 27 members and simultaneously analyzed one unaffected and three affected members by WES. No disease-causing mutations or copy-number variants had been identified by standard sequencing or multiplex ligation-dependent probe amplification (MLPA) of known candidate genes. Linkage and haplotype analyses, instead, mapped the disease locus to an 8.8-Mb region on chromosome 1p31, with a maximum multipoint LOD score of 3.6. The authors selected 11 candidate genes from the critical region using several criteria. Specifically, the candidates were associated or similar to genes associated with other types of inherited retinal degeneration, were included in the sensory cilium proteome or EyeSAGE data, or were highly expressed in the retina. Standard sequencing in two affected members and validation of variants found in both of them using the remainder of the family led to the identification of only one variant, located in *RPE65*. Meanwhile, WES in three affected individuals allowed the identification of 3,437 new variants, reduced to 1,373 after excluding variants that were synonymous or also found in one control DNA. Only three variants remained when restricting to those located in the linkage

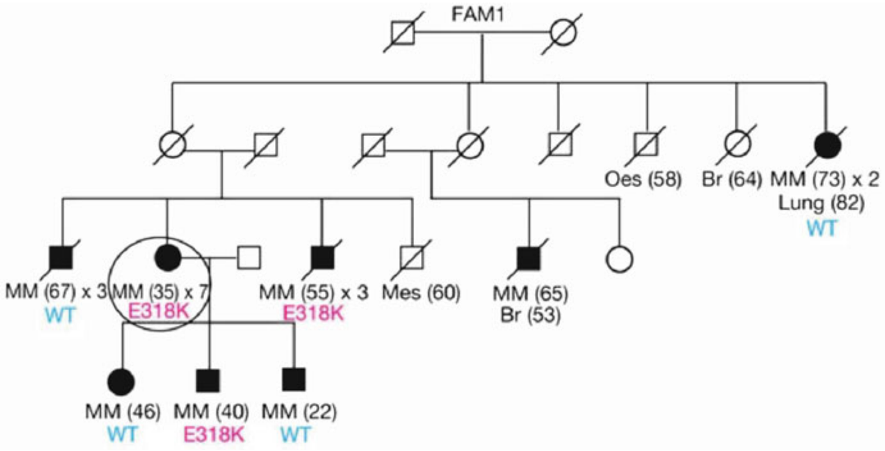

Fig. 3.1 Family 1 showing segregation of the *MITF* E318K variant in some, but not all, affected individuals (Yokoyama et al. 2011). The *circled* individual is the one in which the variant was identified by whole genome sequencing

region, and, of those three variants, the only one present in all three affected individuals was the same *RPE65* variant identified with the concurrent linkage-based approach. Evaluation of 12 Irish patients with a range of other inherited retinal degenerations revealed that one patient, as well as his two affected daughters, initially diagnosed with choroideremia but without a mutation in the *CHM* gene, had the *RPE65* mutation on the same haplotype as the extended adRP family. Combined linkage analysis of the two pedigrees yielded a maximum two-point LOD score of 5.3 at 0% recombination from the mutation. The authors commented that mutations in *RPE65* are a known cause of recessive RP and Leber congenital amaurosis but had never been associated with dominant disease. The less severe phenotype with reduced penetrance observed in the families studied by Bowne et al. (2011) was consistent with one mutant allele rather than two. The authors thus warned that carriers of “recessive” missense mutations in *RPE65* should be evaluated for subtle signs of disease. They also suggested that, given the co-occurrence of choroidal disease in the large adRP family and the diagnosis of choroideremia in the smaller family, mutations in *RPE65* may be the cause of choroideremia in families in which the typical X-linked gene, *CHM*, has been excluded.

While many studies have performed linkage analysis prior or in parallel to the WES/WGS experiments, as exemplified in the previous paper, others have used it to confirm co-segregation of the disease phenotype and variants usually identified in a small number of cases prior to further genetic studies and functional characterization. Yokoyama et al. (2011) looked for germ line mutations predisposing to melanoma starting from whole genome sequencing of a single individual in a family with eight affected relatives in three generations (Fig. 3.1) (Yokoyama et al. 2011). From 410 novel variants thus identified, a variant in *MITF*, a gene known for being somatically amplified or mutated in a subset of melanomas, was found to be present in three out of seven cases tested in the proband's family. Testing of additional patients from families with multiple melanoma cases eventually confirmed the presence of the *MITF* E318K variant in 31 unrelated cases with at least one first- or second-degree relative diagnosed with melanoma. A formal linkage analysis of melanoma with E318K was performed under a dominant model with reduced penetrance and a 5% phenocopy rate and produced a maximum LOD score of 2.7, a result consistent with E318K being an intermediate risk variant. Finally, the authors confirmed the role of E318K in melanoma by means of case–control association studies as well as expression profiling and analysis.

3.3 WES/WGS with Inconclusive Linkage Data

Many linkage studies have resulted in the identification of candidate regions that, however, have not led to the discovery of a specific disease gene. While some of these failures may be explained by false-positive findings (Ioannidis 2005), in other cases the size of the candidate region(s) may simply have prevented its full sequencing and therefore the disease gene identification. With the advent of next-generation sequencing (NGS) technologies, it is increasingly becoming more cost-effective to sequence the whole exome rather than a few target regions that may be relatively large in physical size and number of positional candidate genes contained.

A recent example of the use of this strategy reported by Louis-Dit-Picard et al. (2012) has led to the identification of *KLHL3* mutations in familial hyperkalemic hypertension (FHHt) (Louis-Dit-Picard et al. 2012). SNP-based linkage analysis in one informative family with five affected and seven unaffected individuals indicated six suggestive linkage regions (max LOD = 1.8), spanning a total of 35.6 Mb and containing 325 protein-coding genes. Given the number of positional candidate genes, the authors performed WES of one unaffected and three affected family members. A missense mutation in *KLHL3* was identified in one of the linkage regions on chromosome 5q31, and the same region was reported to be linked in a second family by microsatellite analysis (max LOD = 7.3). WES of three members of the second family also identified one missense mutation in *KLHL3*. Direct sequencing identified nonsynonymous *KLHL3* mutations in 11 out of 14 additional FHHt patients, including heterozygous as well as homozygous cases.

Sobreira et al. (2010) used WGS in a single individual combined with linkage analysis to identify the gene mutated in metachondromatosis (MC), another autosomal dominant disorder (Sobreira et al. 2010). Linkage analysis in seven members of a family segregating MC had identified six regions with positive LOD scores covering a total of 42 Mb, of which one reached the maximum possible in the small pedigree (7p14.1, LOD = 2.5) and two others were compatible with the presence of a single non-penetrant carrier (8q24.1 and 12q23, LOD = 1.8). Following WGS of a single proband, no variants unique to her and with a high likelihood of functional significance were found in five of these regions. However, one such variant was located in the 12q23 candidate region and was shown to be present in all affected individuals as well as the hypothetical non-penetrant carrier who, on more careful examination, showed symptoms of the disease, as well as did her daughter who had not previously been examined. This result clearly points out a possible explanation for the negative results of linkage studies that have been followed up by sequencing only the regions of maximum LOD scores as well as the importance, never overstated, of careful phenotyping. When linkage is inconclusive, like it was in this case, and more than one candidate region exists, it is now more efficient to perform a WES or even a WGS experiment rather than sequencing several candidate regions using more traditional approaches. Interestingly in this case, although the authors performed a WGS study, sequencing of the exome only would have been just as fruitful.

3.4 Homozygosity Mapping and WES/WGS Studies

Homozygosity mapping is a powerful approach for disease gene mapping in cases of rare recessive disease observed in consanguineous families (Lander and Botstein 1987). Homozygosity mapping is a variation on linkage analysis that exploits the fact that in rare autosomal recessive disorders, affected individuals, especially those born from consanguineous parents, are expected to be homozygous for alleles identical by descent (IBD) at the disease locus and at the marker loci tightly linked to it, a condition sometimes referred to as autozygosity. A search for regions of linkage can thus be achieved by looking for regions of IBD homozygosity in a few affected individuals. This approach has been applied successfully to several rare autosomal recessive disorders and is now being further revamped by pairing it with results of WES or WGS experiments.

A good example of this approach is the study of Wang et al. (2011), aimed at identifying novel disease alleles or genes involved in Leber congenital amaurosis (LCA) by combining genetic mapping with WES (Wang et al. 2011). LCA is a genetically heterogeneous eye dystrophy that most often presents as a recessive disease. Standard Sanger sequencing of the coding exons from all 15 known LCA disease genes in one affected member from a consanguineous family from Saudi Arabia had failed to find the causative homozygous mutation. The authors then performed homozygosity mapping by genotyping three affected members using the Illumina 370 K SNP array and identified a single novel region of homozygosity spanning 11.2 Mb on chromosome 15 shared by all three affected members. Due to the high gene density in this region, direct Sanger sequencing of all coding exons was unfeasible. By WES of a single affected individual, they found a total of 370,000 SNPs and in/dels. After filtering out common variants and variants that did not affect protein-coding or splicing regions, they were left with 352 candidate variants. The only homozygous missense change in the critical region was located in *BBS4*, a gene known to cause Bardet-Biedl syndrome (BBS). BBS is a rare human genetic disorder that, similarly to LCA, presents ocular phenotypes as a common clinical feature. The authors confirmed the presence of the mutation with Sanger sequencing and that it segregated with the disorder in the family by direct genotyping of all the other members. Moreover, they excluded the presence of this variant in 200 normal matching controls, including 96 from Saudi Arabia, and confirmed its pathological role in a zebrafish model.

Using a similar approach, Schrader et al. (2011) investigated an extended family that presented with autosomal recessive spondyloepiphyseal dysplasia (SED), retinitis pigmentosa (RP), and a high incidence of corneal abnormalities among affected individuals (Schrader et al. 2011). Given the geographical isolate from which the family originated, the known consanguinity, and the autosomal recessive inheritance pattern of the disease, the authors hypothesized that the causative mutation would be novel and would lie within an extended block of linkage that was homozygous in the affected individuals and heterozygous in the unaffected obligate carriers. For this reason, they performed WES in three affected individuals and one unaffected obligate carrier from the family and in parallel applied SNP chip genotyping to the same individuals to rule out homozygous microdeletions and to identify blocks of linkage surrounding candidate novel variants. Among the variants detected by WES, only two uncommon ones were homozygous in all three affected individuals and heterozygous in the obligate carrier: a nonsynonymous variant in *RPL3L* and a 6-bp deletion in *GNPTG*. These variants were validated by Sanger sequencing and found to co-segregate with the disease in the other 14 family members. Furthermore, both variants were located within a 3.5-Mb region of linkage defined by homozygosity in affected individuals, containing 202 UCSC genes. The authors focused their analysis on the mutation in *GNPTG,* a gene associated with mucolipidosis type IIIγ (MLIIIg), an autosomal recessive lysosomal storage disorder with a broad phenotypic spectrum including progressive destruction of the hip joint, increased lysosomal enzyme levels in serum, and reduced lysosomal enzyme levels in cultured fibroblasts. Elevated lysosomal enzyme activity was confirmed in the serum of affected individuals, and histochemical analysis of a section of the femoral head of one member of the family revealed microvesicular changes in the chondrocytes. Thus, their approach eventually led to a molecular diagnosis of MLIIIg and to a further broadening of the phenotypic spectrum of MLIII. These authors compared the traditional linkage mapping, homozygosity mapping, and whole exome sequencing approaches and concluded that the latter should be sufficient to identify causal mutations in most Mendelian disorders. However, they did recommend SNP array genotyping in at least one individual to rule out homozygous deletions and duplications that could be missed otherwise.

The study of Puffenberger et al. (2012) in the Amish and Mennonite populations of Pennsylvania represents perhaps the best example of the power of the combined homozygosity mapping/WES approach (Puffenberger et al. 2012). Taking advantage of the characteristics of these populations, including relative isolation and high inbreeding, the same authors had previously identified the loci for 28 genetic disorders by homozygosity mapping. For 11 of these, however, the corresponding gene could not be found, and the authors cited the large size of the candidate regions and large number of genes there contained as the main obstacles to achieving this goal. The authors looked at seven such diseases where gene mapping had been achieved by SNP genotyping using either a 10K or 50K SNP microarray. In six cases, only one candidate region had been identified using either two or more affected individuals from a single family or multiple cases from different families; in the remaining case, analysis of two affected siblings, their parents, and six unaffected siblings resulted in the identification of 12 candidate genomic regions each greater than 5 Mb. Even when a single genomic region was consistent with linkage, the average size of the candidate regions was 4.4 Mb (range 1.6–8.4 Mb), and the average number of genes included in them was 79 (range 22–187). Sequencing a number of candidate genes included between 2 and 45 for each condition failed to identify the causative variants. In contrast, WES of a number of patient samples included between one and five (for a total of 15 cases for all disorders) and subsequent filtering of candidate variants led to the identification of a single causal mutation in all seven diseases, five of which located in genes that had not previously been associated with these conditions. Criteria for disease variant identification included homozygosity in the affected patients, localization in the regions of linkage, and absence from dbSNP 129 and 1000 Genomes Project. All putative disease variants were confirmed by Sanger sequencing in the cases and their available relatives; their frequency in the population was further evaluated in more than 400 chromosomes, and no homozygous controls were identified. In some instances, the presence of the same mutation was confirmed in independent cases with the same phenotype, or pathogenicity was supported by high PolyPhen2 scores. Finally, *in vitro* studies supported the causal relationship between some of the candidate variants and the respective disease phenotype.

Interestingly, these authors noted that when multiple cases were available, the use of WES coupled with the assumption of mutation homozygosity in the patients but not in unaffected individuals would have been sufficient to identify the disease gene mutations even in the absence of mapping data. In fact, in each of these cases, only one variant was identified that satisfied these conditions. Even when only a single case was sequenced, the number of potentially pathogenic homozygous variants was relatively small, perhaps surprisingly given the high inbreeding coefficients of these populations, and only six variants were not homozygous in unaffected controls. In conclusion, the authors suggested that a strategy based on WES and a search for homozygous or compound heterozygous novel variants in the same gene in multiple affected individuals has a high chance of being successful even in outbred populations.

However, a cautionary note on the use of WES comes from the study of Bloch-Zupan et al. (2011) of two first-degree cousins affected with major dental developmental defects (Bloch-Zupan et al. 2011). Because of the high consanguinity in the family, the authors used homozygosity mapping to identify a critical region for the disease gene located on 6q27-ter and spanning 3 Mb. Sequencing of two candidate genes in the critical region led to the identification of a splice site mutation in the *SMOC2* gene that was present in the homozygous state in the two children and in the heterozygous state in the children's carrier parents. To confirm that no other mutations were present in the children that may explain the phenotype, Bloch-Zupan et al. performed WES in one of the two patients. Interestingly, they found out that 6.6 Kb of the 3-Mb critical region identified by homozygosity mapping were not sufficiently covered by the WES data, and specifically the mutation in *SMOC2* identified by traditional sequencing was not detected by WES. Analysis of the genomic region containing the mutation showed it to be GC rich, and independent sequencing experiments from other projects confirmed the deficit in sequence coverage. The authors concluded that had they only applied the exome-capture approach, they would have missed the causative mutation in their patients.

3.5 Linkage and WES/WGS in Quantitative Trait Analysis

Integration of linkage analysis and sequencing studies can also be used successfully for the identification of the molecular basis underlying a quantitative trait locus (QTL), as exemplified by the study of Bowden et al. (2010) on adiponectin plasma levels (Bowden et al. 2010). In this case, variance-component linkage analysis, a popular approach for QTL mapping, had identified a strong linkage signal in a single genomic region (3q, LOD = 8.02). The linkage critical region contained an ideal candidate for variation in adiponectin plasma level, the adiponectin protein-coding gene *ADIPOQ*. However, association to common variants in this gene did not explain the linkage signal. The authors cleverly used the linkage results to select individuals for sequencing by prioritizing families with a higher individual LOD score in the critical region. WES of three individuals with values of adiponectin plasma level in the tails of its distribution (one high, two low) from two of these families led to the identification of a single variant not previously reported and present in the two low-adiponectin samples. Through conventional sequencing and additional genotyping, the same rare variant was shown to co-segregate with the plasma adiponectin trait in the linkage families and to account for most of the 3q linkage signal. While this study may be considered just a proof of principle given the presence of a very strong candidate gene in the linkage critical region, it showed that the combination of QTL linkage mapping and sequencing of individuals with extreme values of the quantitative trait is a potentially valuable approach for the identification of rare variants, as it has been recently advocated (Cirulli and Goldstein 2010).

3.6 Linkage Analysis as an Aid in Designing WES/WGS Experiments

Bowden et al. (2010) have shown that results from linkage analysis can be utilized in WES/WGS projects to optimize family selection (Bowden et al. 2010). The GAW17 dataset provided an opportunity to investigate the efficacy and cost efficiency of various strategies for next-generation sequencing sample selection as a follow-up to linkage analysis (http://www.gaworkshop.org). The GAW17 dataset included genome-wide genotype data as well as exome sequencing for approximately 3,000 genes for eight simulated pedigrees. In addition, risk factors including age, smoking status, and three quantitative trait variables were provided for each individual. Allen-Brady et al. (2011) performed linkage analyses on these families and compared nine approaches for selecting subjects for subsequent partial exome sequencing (Allen-Brady et al. 2011). For this study, the authors split the original 8 pedigrees into 23 smaller pedigrees and used 10 of the 200 replicate datasets provided by the GAW17 organizers. Using the results from a logistic regression model which incorporated the five risk factors, Allen-Brady et al. (2011) classified all individuals as either high-covariate subjects whose nongenetic risk factors are highly predictive of their affection status or low-covariate subjects whose nongenetic risk factors are poorly predictive of their affection status. They found that selecting for exome sequencing all affected individuals classified as low-covariate and possessing a linked haplotype identified in the linkage analysis was the most reliable strategy across both recessive and dominant models. Furthermore, selecting the youngest affected individuals may provide a satisfactory alternative in cases where the major nongenetic risk factors are unknown.

Starting from the GAW17 pedigrees, Cai et al. (2011) defined as high risk those with at least 15 total meioses between case subjects and a statistical excess of disease ($p < 0.01$) over all 200 replicates, thus identifying 18 pedigrees (Cai et al. 2011). They then performed a linkage analysis using the shared genomic segment (SGS) method (Thomas et al. 2008), modified in order to examine sharing among all pairs of cases instead of all subjects, and assessed the test statistic against an empirical distribution. Following this approach, they successfully identified at least one region containing one true causal variant in 13 out of the 18 high-risk pedigrees. Additional causative genes would have been identified at lower significance thresholds ($p \leq 0.01$), but this would also have increased false-positive findings. The inability to detect the other rare causative variants was ascribed to the small sample size and high heterogeneity. Of note, this method considered only pairs of relatives and did not take into account the specific relationships between them. The authors claimed that when they incorporated the specific relationships into the analysis, they did not see substantial improvement in the results.

Gagnon et al. (2011) analyzed the GAW17 data to select which families should be sequenced in order to identify rare variants that have large effects on quantitative trait variance (Gagnon et al. 2011). They hypothesized that rare functional variants segregating with a quantitative phenotype are more likely to be present in families with more quantitative trait loci (QTLs) than the other families. For this reason, they estimated the mean number of QTLs in each family by segregation analysis assuming an oligogenic linear model and selected one family with more QTLs than the average. They then tested this family for linkage using a variance-component oligogenic approach. Sequencing data from regions surrounding loci with at least modest evidence of linkage (LOD ≥ 0.6) were investigated for the presence of rare functional variants, and the variants thus detected were analyzed for association to the quantitative traits. By this approach, they identified one region with a maximum LOD of 5.3 ($p = 4 \times 10^{-7}$) for one trait and two regions with maximum LOD of 2.02 ($p = 0.001$) for another trait for a total of 216 and 85 variants that were thus tested for association with the same traits in all the families. They correctly identified two rare functional variants, including one private to the family selected for sequencing. Both variants were located in regions with a combined LOD score in all families greater than 4. All other variants identified in the family selected for linkage analysis were false positives, and all had LOD scores below 2, confirming once again the importance of linkage evidence in discovering the actual causative variants. The authors claimed that prioritizing the sequencing of carefully selected extended families is a simple and cost-efficient design strategy to identify rare functional variants that explain a significant proportion of the trait variance, especially for variants that are unlikely to segregate in more than a few families. However, they noted that the use of large, multigenerational families remains crucial, and other complementary designs are still needed to further decrease type I error, including parallel analysis of large samples of unrelated individuals.

Other research utilizing the GAW17 dataset focused on the potential of linkage studies to guide deep sequencing efforts by narrowing the search space to genomic regions under linkage peaks. Choi et al. (2011) compared the effectiveness of two mapping strategies: (1) performing association tests which adjust for familial relationships on variants identified by whole exome sequencing and (2) performing a linkage analysis followed by targeted sequencing of regions beneath the linkage peaks and family-based association on variants identified in those regions (Choi et al. 2011). They found that both mapping strategies demonstrate a limited ability to detect association for variants of small effect sizes. In addition, both strategies only found the same two loci with a reasonable amount of power (>70%). However, the linkage-guided strategy on average required sequencing of only 2.5% of the whole exome and found 52% of the associated loci identified by the whole exome sequencing strategy. Choi et al. concluded that while the whole exome sequencing strategy appears more powerful, targeted sequencing under linkage peaks still offers a viable and cost-effective alternative.

3.7 New Methods of Linkage Analysis and WES/WGS Studies

One of the challenges of linkage analysis has always been the analysis of large pedigrees, which, however, can also provide very valuable information. In place of traditional linkage analysis as a filter for whole exome sequence data, some groups have attempted to find equally effective strategies that are less computationally intense. Markello et al. (2012) advocated using high-density genotyping panels and Boolean logic for recombination mapping efforts (Markello et al. 2012). High-density genotyping panels provide a relatively cost-effective option that covers introns as well as intergenic regions as compared to a strategy that extracts SNPs directly from WES. However, double-crossover events in between contiguous informative markers are problematic for this approach. Markello et al. (2012) showed that the risk of double-crossover events between informative SNPs was extremely low through simulation and a literature search to identify the smallest reported interval containing a double crossover. In addition, these authors compared their recombination mapping method to a traditional multipoint linkage analysis utilizing microsatellite markers and demonstrated that the identified regions were identical. Finally, they incorporated the method into a filtering scheme on variants identified by whole exome sequencing of a single pedigree and identified two potential disease-causing mutations in a candidate gene for progressive myoclonic epilepsy type 3.

Marchani and Wijsman (2011) proposed a method to test for linkage that enjoys computational advantages because of how it records and groups the inheritance patterns (Marchani and Wijsman 2011). Their method can employ Markov chain Monte Carlo (MCMC) techniques to handle large pedigrees and can also be applied to common disorders where Mendelian inheritance is not strictly followed and genetic heterogeneity is present. Visualization of IBD sharing allows the investigator to observe which affected pedigree members share a genetic segment within a region tied to a linkage signal. This knowledge allows efficient selection of individuals for deep sequencing. The strategy advocated by Marchani and Wijsman (2011) is to select the most distantly related affected individuals who share such a DNA segment.

Other research has focused on the utility of gleaning variants directly from next-generation sequencing for use in linkage studies. Smith et al. (2011) investigated the efficacy of using genotypes generated from WES as a surrogate for array-based genotypes in linkage studies (Smith et al. 2011). Limitations of the WES approach include coverage gaps in non-exonic regions, higher genotyping error rates, and markers with lower heterozygosity. Smith et al. (2011) performed linkage analyses on three pedigrees with different Mendelian neurological disorders employing both array-based markers and HapMap phase II SNP genotypes derived from WES. They found (1) a substantial number of WES-derived SNPs resided outside of coding regions due to a technical artifact of the sequencing method, (2) almost a 100% concordance rate for genotypes derived from either of the two methods, signifying an acceptable error rate for WES-derived SNPs, and (3) the resulting LOD scores for the analyses using genotypes derived from WES closely resembled those for the analyses using genotypes acquired by array-based technology at the positions of linkage peaks. Smith et al. (2011) concluded that, while SNP arrays are preferable for linkage studies due to better coverage and marker informativeness, generating genotypes for linkage studies directly from WES data is a viable option.

3.8 Conclusions

The advent of NGS technologies and the ability to sequence whole exomes or genomes have generated a new interest in the analysis of family data and thus in genetic linkage analysis (Bailey-Wilson and Wilson 2011). Linkage analysis is ideal for identifying the location of rare disease-causing variants, such as those that are the object of analysis in most sequencing studies. Candidate loci identified by linkage studies can now be examined more extensively than ever before, leading to a new wave of gene discoveries, particularly for Mendelian disorders (Ng et al. 2010).

Whether the same success will occur in the analysis of complex traits remains to be demonstrated. Nonetheless, it is especially important that all available approaches are considered when tackling a difficult question such as the identification of the genetic basis of complex disease, and we recommend that linkage analysis should be considered whenever families are available to investigators.

References

Allen-Brady K, Farnham J, Cannon-Albright L. Strategies for selection of subjects for sequencing after detection of a linkage peak. BMC Proc. 2011;5 Suppl 9:S77.

Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. Hum Hered. 2011;72(4):228–36.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745–55.

Bloch-Zupan A, Jamet X, Etard C, Laugel V, Muller J, Geoffroy V, Strauss JP, Pelletier V, Marion V, Poch O, Strahle U, Stoetzel C, Dollfus H. Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. Am J Hum Genet. 2011;89(6):773–81.

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl):228–37.

Bowden DW, An SS, Palmer ND, Brown WM, Norris JM, Haffner SM, Hawkins GA, Guo X, Rotter JI, Chen YD, Wagenknecht LE, Langefeld CD. Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS family study. Hum Mol Genet. 2010;19(20):4112–20.

Bowne SJ, Humphries MM, Sullivan LS, Kenna PF, Tam LC, Kiang AS, Campbell M, Weinstock GM, Koboldt DC, Ding L, Fulton RS, Sodergren EJ, Allman D, Millington-Ward S, Palfi A, McKee A, Blanton SH, Slifer S, Konidari I, Farrar GJ, Daiger SP, Humphries P. A dominant mutation in RPE65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. Eur J Hum Genet. 2011;19(10):1074–81.

Cai Z, Knight S, Thomas A, Camp NJ. Pairwise shared genomic segment analysis in high-risk pedigrees: application to genetic analysis workshop 17 exome-sequencing SNP data. BMC Proc. 2011;5 Suppl 9:S9.

Choi SH, Liu C, Dupuis J, Logue MW, Jun G. Using linkage analysis of large pedigrees to guide association analyses. BMC Proc. 2011;5 Suppl 9:S79.

Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.

Gagnon F, Roslin NM, Lemire M. Successful identification of rare variants using oligogenic segregation analysis as a prioritizing tool for whole-exome sequencing studies. BMC Proc. 2011;5 Suppl 9:S11.

Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.

Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236(4808):1567–70.

Louis-Dit-Picard H, Barc J, Trujillano D, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O, Beaurain G, Bonnefond A, Sand O, Simian C, Vidal-Petiot E, Soukaseum C, Mandet C, Broux F, Chabre O, Delahousse M, Esnault V, Fiquet B, Houillier P, Bagnis CI, Koenig J, Konrad M, Landais P, Mourani C, Niaudet P, Probst V, Thauvin C, Unwin RJ, Soroka SD, Ehret G, Ossowski S, Caulfield M, Bruneval P, Estivill X, Froguel P, Hadchouel J, Schott JJ, Jeunemaitre X. KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Nat Genet. 2012;44(5):609.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.

Marchani EE, Wijsman EM. Estimation and visualization of identity-by-descent within pedigrees simplifies interpretation of complex trait analysis. Hum Hered. 2011;72(4):289–97.

Markello TC, Han T, Carlson-Donohoe H, Ahaghotu C, Harper U, Jones M, Chandrasekharappa S, Anikster Y, Adams DR, Gahl WA, Boerkoel CF, Program NCS. Recombination mapping using Boolean logic and high-density SNP genotyping for exome sequence filtering. Mol Genet Metab. 2012;105(3):382–9.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010;42(1):30–5.

Puffenberger EG, Jinks RN, Sougnez C, Cibulskis K, Willert RA, Achilly NP, Cassidy RP, Fiorentini CJ, Heiken KF, Lawrence JJ, Mahoney MH, Miller CJ, Nair DT, Politi KA, Worcester KN, Setton RA, Dipiazza R, Sherman EA, Eastman JT, Francklyn C, Robey-Bond S, Rider NL, Gabriel S, Morton DH, Strauss KA. Genetic mapping and exome sequencing identify variants associated with five novel diseases. PLoS One. 2012;7(1):e28936.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273(5281):1516–17.

Schrader KA, Heravi-Moussavi A, Waters PJ, Senz J, Whelan J, Ha G, Eydoux P, Nielsen T, Gallagher B, Oloumi A, Boyd N, Fernandez BA, Young TL, Jones SJ, Hirst M, Shah SP, Marra MA, Green J, Huntsman DG. Using next-generation sequencing for the diagnosis of rare disorders: a family with retinitis pigmentosa and skeletal abnormalities. J Pathol. 2011;225(1):12–8.

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol. 2011;12(9):R85.

Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, Gumbs CE, Pevsner J, Thomas G, Valle D, Hoover-Fong JE, Goldstein DB. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. 2010;6(6):e1000991.

Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. Ann Hum Genet. 2008;72(Pt 2):279–87.

Wang H, Chen X, Dudinsky L, Patenia C, Chen Y, Li Y, Wei Y, Abboud EB, Al-Rajhi AA, Lewis RA, Lupski JR, Mardon G, Gibbs RA, Perkins BD, Chen R. Exome capture sequencing identifies a novel mutation in BBS4. Mol Vis. 2011;17:3529–40.

Yokoyama S, Woods SL, Boyle GM, Aoude LG, MacGregor S, Zismann V, Gartside M, Cust AE, Haq R, Harland M, Taylor JC, Duffy DL, Holohan K, Dutton-Regester K, Palmer JM, Bonazzi V, Stark MS, Symmons J, Law MH, Schmidt C, Lanagan C, O'Connor L, Holland EA, Schmid H, Maskiell JA, Jetann J, Ferguson M, Jenkins MA, Kefford RF, Giles GG, Armstrong BK, Aitken JF, Hopper JL, Whiteman DC, Pharoah PD, Easton DF, Dunning AM, Newton-Bishop JA, Montgomery GW, Martin NG, Mann GJ, Bishop DT, Tsao H, Trent JM, Fisher DE, Hayward NK, Brown KM. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. Nature. 2011;480(7375):99–103.

Zeggini E. Next-generation association studies for complex traits. Nat Genet. 2011;43(4):287–8.

Chapter 4
From Family Study to Population Study: A History of Genetic Mapping for Nasopharyngeal Carcinoma (NPC)

Timothy J. Jorgensen, Hai-De Qin, and Yin Yao Shugart

Abstract Nasopharyngeal carcinoma (NPC) has a unique global distribution pattern – Southeast Asia and some other localized regions of the eastern hemisphere – that suggests risk is largely driven by a combination of environmental exposures and specific genetic factors. Earlier linkage analysis has implicated loci in the human leukocyte antigen (HLA) gene region, thus suggesting a role for immunological mechanisms in NPC resistance. Nevertheless, the implications of the *HLA* associations remain enigmatic. More recent association studies have sought to advance our understanding of the genes important to NPC risk. Reviewed here are recent epidemiologic studies that have addressed the genetics of NPC risk, and the implications of their collective findings are discussed. The primary focus is on the latest candidate-gene association studies (CGAS) and genome-wide association studies (GWAS), and attempts are made to harmonize their findings and resolve discrepancies. Taken together, the studies support the importance of the *HLA* loci, but also implicate non-HLA genes both inside and outside the *HLA* region, and suggest that the mechanisms of NPC risk go beyond immunology. Finally, recommendations

T.J. Jorgensen (✉)
Department of Radiation Medicine, Georgetown University Medical Center,
3970 Reservoir Road, NW, TRB Room E220, Washington, DC 20057, USA
e-mail: tjorge01@georgetown.edu

H.-D. Qin
Unit of Statistical Genomics, Division of Intramural Research Programs,
National Institute of Mental Health (NIMH)/NIH, Building 35; Room 3A-1006,
Bethesda, MD, USA
e-mail: qinh2@mail.nih.gov

Y.Y. Shugart
Unit of Statistical Genomics, Intramural Research, Program, National Institute of Mental Health,
35 Convent Drive, Room 3A1000, Bethesda, MD, 20892-3719, USA
e-mail: kay1yao@mail.nih.gov

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_4, © Springer Science+Business Media Dordrecht 2012

are made to coordinate future CGAS and GWAS to maximize their information content and make best use of the limited number of available NPC study populations.

Keywords Nasopharyngeal carcinoma • Candidate-gene study • Genome-wide association study • *HLA*

4.1 Introduction

Current understanding of cancer etiology suggests both genetic factors and environmental exposures play important roles in causation. A major goal of cancer research has been to characterize the interplay between these genetic and environmental causes. In this regard, nasopharyngeal carcinoma (NPC) is of great interest because its unique global distribution pattern suggests that risk is largely dependent upon a combination of specific genetic factors and distinct environmental exposures. For this reason, NPC can be considered a paradigm for cancer genetics (Simons 2011) and provides a unique opportunity to inform our understanding of the mechanisms of human carcinogenesis.

Regarding environmental risk factors, the strongest associations have been made with Epstein-Barr virus (EBV) infection and consumption of salt-preserved fish. Much weaker associations have been made with tobacco smoke and alcohol. The epidemiological literature on these environmental NPC risk factors is vast, and several earlier reviews comprehensively summarize the findings (Brennan 2006; Chang and Adami 2006; Gallicchio et al. 2006; Jeyakumar et al. 2006; Wei et al. 2010a; Cao et al. 2011).

There is nearly 40 years of evidence suggesting that genetic factors are also major drivers of NPC risk. Immigrants from high- to low-risk NPC areas maintain their high NPC risk (Parkin and Iscovich 1997). Also, family, twin, and segregations studies support genetic factors as strong determinants of NPC risk (Gajwani et al. 1980; Zeng and Jia 2002; Jia et al. 2005; Ng et al. 2009). More specifically, there are multiple reported associations between NPC and loci linked to the regions of the genome where human leukocyte antigen (HLA) genes reside, yet the implications of the *HLA* allelic associations remain enigmatic. And the importance of other associated loci both within and outside the *HLA* regions has not been thoroughly investigated. There also is limited understanding of how the major environmental risk factors interact with the genotypes.

In this review, we summarize the more recent genetic epidemiology reports (i.e., last 15 years) regarding NPC risk. We also attempt to identify patterns of evidence within and among studies that bolster the findings. Further, we discuss the implications of the genetic aspects in relation to the environmental risk factors. We concentrate mainly on the latest candidate-gene association studies (CGAS) and genome-wide association studies (GWAS) and attempt to harmonize their findings and provide potential justifications for any discrepancies. Finally, we suggest new avenues for future investigations.

4.2 The Working Model

Virtually all NPC tumors express EBV proteins, while normal nasopharyngeal tissues do not. And the tendency to reactivate latent EBV virus is highly correlated with NPC risk – so much so, that measurement of EBV reactivation is often used as an early cancer biomarker in NPC endemic regions (Li et al. 2010). Yet, EBV infection is highly prevalent and pandemic, while NPC incidence is low in most parts of the world. Nevertheless, in Southeast Asia and some other localized regions of the eastern hemisphere, NPC incidence is high and tends to be clustered in families. Thus, EBV infection seems to be a necessary but insufficient component of the NPC causal mechanism. This has led to the proposition that certain individuals carry genetic variants that predispose them to the carcinogenic transforming potential of EBV, and that these variants are relatively common among the people in NPC endemic regions.

This is a useful working model of NPC carcinogenesis since the molecular mechanisms of EBV reproduction and infection are well known, and the mechanisms of EBV carcinogenic transformation have been intensively investigated in the laboratory (Rowe 1999; Hatzivassiliou and Mosialos 2002; Liu et al. 2006; Martin and Gutkind 2008; Pang et al. 2009). Thus, this model identifies a number of specific host genes that may interact with EBV, and these genes constitute promising candidates for investigation in candidate-gene association studies (CGAS).

Genes with potential relevance to NPC and their biochemical functions were the subject of a review by Chou and coworkers (Chou et al. 2008). These genes can be clustered into biochemical pathways with specific functions, and this has allowed a pathway-based approach to both define the universe of potentially associated genes and facilitate the analytical process (Jorgensen et al. 2009; Thomas et al. 2009a). EBV-related host genes have been the favored genes for interrogation in most of the more recent CGAS. We will, therefore, primarily focus on these candidate genes here but will also consider genes from other pathways potentially related to NPC.

4.3 Candidate-Gene Association Studies

The CGAS approach has several advantages, the biggest one being that having a strong prior probability reduces the number of variant alleles that must be assessed and, thereby, preserves statistical power that would otherwise be reduced due to the statistical corrections needed to account for multiple comparisons. The increased power is particularly important for interrogations of smaller populations with lower case numbers. Below we review, by metabolic pathway, recent CGAS investigations of NPC (<15 years) that have at least 45 cases and were published in English.

4.3.1 Apoptosis and Cell Cycle Arrest Pathways

Apoptosis – a programmed cell death that eliminates transformed cells – is thought to be downregulated in many types of tumors, including NPC. The genes that regulate apoptosis often overlap with the regulatory genes for cell cycle arrest – a protective response to DNA damage that allows cells time to repair damage before cell replication proceeds – since apoptosis is often a consequence of faulty arrest. EBV is known to inhibit apoptosis by a mechanism that is thought to involve expression of viral transforming protein LMP1 (Xiong et al. 2004; Grimm et al. 2005; Zheng et al. 2007a; Chew et al. 2010), and also to concurrently inhibit cell cycle arrest (Pokrovskaja et al. 1999; O'Nions and Allday 2003). Therefore, although apoptosis and cell cycle arrest represent very different functions, promoting either cell death or survival, respectively, the genes for each pathway will be discussed here collectively.

The central player of cell cycle arrest and apoptosis functions is the *TP53* gene, which codes for the p53 protein. This protein governs both DNA damage-dependent cell cycle arrest and apoptosis. EBV nuclear antigen 3C is thought to modulate cellular apoptosis by inhibiting transcription of p53 (Saha et al. 2009; Yi et al. 2009), and thus may contribute to carcinogenesis. Five (Tsai et al. 2002b; Tiwawech et al. 2003; Sousa et al. 2006; Hadhri-Guiga et al. 2007; Xiao et al. 2010) of the eight apoptosis and cell cycle arrest CGAS (Deng et al. 2002; Tsai et al. 2002a, b; Tiwawech et al. 2003; Cao et al. 2006; Sousa et al. 2006; Hadhri-Guiga et al. 2007; Xiao et al. 2010) looked at p53, and four of these reported significant associations between *TP53* alleles and NPC (Tsai et al. 2002b; Sousa et al. 2006; Hadhri-Guiga et al. 2007; Xiao et al. 2010). Four studies that specifically looked at a nonsynonymous SNP in codon 72 (Tsai et al. 2002b; Tiwawech et al. 2003; Sousa et al. 2006; Hadhri-Guiga et al. 2007) were included in a meta-analysis of codon 72 and NPC risk (Zhuo et al. 2009b). [A fifth study that we omitted here due to its low case number (i.e., 20 cases) (Yung et al. 1997) was also incorporated into this meta-analysis.] Meta-analysis results indicated significantly elevated risk associated with the codon 72 proline allele relative to the arginine allele ($P < 0.0003$).

There were also significant associations reported for *FAS* (Cao et al. 2010b) and *MDM2* (Xiao et al. 2010) – genes that are important to apoptosis. Taken together with the meta-analysis for *TP53*, an upstream regulator of apoptosis, the CGAS reports support a role for DNA damage-induced apoptosis, and possibly cell cycle arrest, in NPC risk.

4.3.2 Carcinogen Metabolism and Detoxification Pathways

Studies have shown that cytochrome P450 metabolic pathway is important to resistance to cancer (Rodriguez-Antona et al. 2009), including nasopharyngeal carcinoma (Hou et al. 2007), particularly among EBV seropositive individuals (Hildesheim et al. 2001). It has further been demonstrated that the carcinogenic activity of nitrosamines requires bioactivation by cytochrome P450 2E1 (*CYP2E1*)

(Yang et al. 1990). N-nitrosamines are among the known components of salt-preserved foods and tobacco (Haorah et al. 2001) – both environmental risk factors for NPC. In particular, nitrosamine metabolism-related DNA adducts have been linked to NPC (Dodd et al. 2006). Furthermore, the metabolites of these carcinogens can generate reactive oxygen species (ROS), which in turn produce base damage, single-strand breaks, and double-strand breaks in DNA (Frenkel 1992). For these reasons, *CYP2E1* and other P450 enzymes have been considered prime candidate genes for association with NPC, and a number of studies have focused on the cytochrome P450 genes (Table 4.1). In addition, the glutathione transferase genes, which are important for recycling glutathione – an extremely important intracellular scavenger of ROS – have also been the focus of studies.

There were a total of 11 studies focusing on carcinogen metabolism and detoxification genes (Hildesheim et al. 1995, 1997; Nazar-Stewart et al. 1999; Cheng et al. 2003; Jiang et al. 2004; Tiwawech et al. 2005; Tiwawech et al. 2006; Guo et al. 2008; He et al. 2009; Jia et al. 2009; Guo et al. 2010), but significant associations were only found for GSTM1, *CYP2A6*, and *CYP2E1*. Of the three, the evidence for *CYP2E1* was strongest. One report showed a relatively high overall risk of 2.6 (95%CI = 1.2, 5.7), but there was no interaction with smoking or alcohol consumption (Hildesheim et al. 1997). Another showed elevated risk only among smokers (Jia et al. 2009). Nevertheless, seven different loci within the gene were statistically significantly associated with NPC, with P values ranging from 0.014 to 0.0001 (Jia et al. 2009). Furthermore, the false-positive report probability for six SNPs was <0.015, suggesting that the associations were unlikely to be false.

For the glutathione transferase genes, a meta-analysis of deletion alleles for *GSTM1* and *GSTT1* was conducted (Zhuo et al. 2009a). It included eight studies, but four were of small size or written in a language other than English. So only four of the studies met our criteria for inclusion here. The meta-analysis indicated a significant association only for *GSTM1* (OR = 1.42; 95%CI = 1.21, 1.66).

4.3.3 DNA Repair Pathways

DNA repair processes are known to be dysregulated in NPC tumor cells (Cheung et al. 2006; Dodd et al. 2006; Sckolnick et al. 2006). And EBV has been shown to both promote DNA damage and interfere with its repair (Liu et al. 2004, 2005; Iwakawa et al. 2005; Bailey et al. 2009; Gruhne et al. 2009; Wu et al. 2009). In addition, it is long established that normal DNA repair capacity is important to cancer resistance (Berwick and Vineis 2000). It has also recently been reported that DNA repair genes may affect seroreactivation of EBV (Shen et al. 2011), which is highly correlated with increased NPC risk (Tam and Murray 1990; Ji et al. 2007). Therefore, DNA repair genes represent good candidates for NPC association studies.

There were eight (Cho et al. 2003; Yang et al. 2007, 2008, 2009; Zheng et al. 2007b, 2011; Cao et al. 2006; Qin et al. 2011) studies of DNA repair genes, encompassing a total of 90 different genes. Significant associations were reported

Table 4.1 Nasopharyngeal carcinoma candidate gene association studies by biological pathway

	Study (first author and year)	Ref	Cases	Cont.	Genes studied	Significant gene associations	Odds ratio	95% CI	*P* value
Apoptosis and cell cycle arrest	Cao 2010b	*30*	582	613	FAS, FASL	FAS[a]	1.69	1.21, 2.35	0.002
	Deng 2002	*31*	84	91	CCND1	CCND1	2.46	1.25, 4.86	0.016
	Hadhri-Guiga 2007	*24*	115	83	TP53	TP53	not reported	not reported	0.0307
	Sousa 2006	*25*	107	285	TP53	TP53	2.67	1.21, 5.90	0.012
	Tiwawech 2003	*26*	102	148	TP53	none	na	na	na
	Tsai 2002	*29*	47	119	WAF1/CIP1	none	na	na	na
	Tsai 2002b	*27*	50	59	TP53	TP53	0.33	0.13, 0.85	<0.05
	Xiao 2010	*28*	522	722	MDM2, TP53	MDM2, TP53	2.83; 2.22	2.08, 3.96; 1.58, 3.10	na
Carcinogen metabolism and detoxification	Cheng 2003	*43*	337	317	CYP1A1, GSTM1, GSTT1,GSTP1, NAT2	none	na	na	na
	Guo 2008	*41*	350	622	GSTM1, GSTT1	none	na	na	na
	Guo 2010	*40*	358	629	CYP2E1, GSTP1, NQO1, MPO	none	na	na	na
	He 2009	*45*	239	286	GSTM1	none	na	na	na
	Hildesheim 1995	*47*	50	50	CYP2E1	none	na	na	na
	Hildesheim 1997	*46*	364	320	CYP2E1	CYP2E1[b]	2.6	1.2, 5.7	na
	Jia 2009	*44*	755	755	CYP2E1	CYP2E1 among smokers (7 loci)[c]	1.88 to 2.99 for smokers	na	0.0001–0.0140
	Jiang 2004	*49*	472	709	CYP2A13	none	na	na	na
	Nazar-Stewart 1999	*48*	83	114	GSTM1	GSTM1[d]	1.9	1.0, 3.3	0.05
	Tiwawech 2005a	*42*	78	145	GSTM1	none[e]	na	na	na
	Tiwawech 2006	*50*	74	137	CYP2A6	CYP2A6	2.37	1.27, 4.46	<0.01
Cell adhesion	BenNasr 2010	*89*	162	140	CDH1	CDH1	2.02	1.20, 3.40	0.008
	Xu 2010	*88*	444	464	DC-SIGN	DC-SIGN[f]	2.10	1.23, 3.59	0.006

Cytokines and growth factors	Gao 2008	*81*	173	206	EGF, EGFR	none	na	na	na
	Nasr 2008	*80*	163	169	VEGF	VEGF	1.4	not reported	0.03
	Wang 2010	*79*	156	161	VEGF	VEGF	1.65	1.05, 2.58	0.029
	Wei 2007	*82*	108	120	TGF-beta1	TGF-beta1 (two loci)	1.63; 1.70	1.13, 2.39; 1.17, 2.46	0.009, 0.006
DNA methylation	Cao 2010a	*102*	529	577	MTHFR	MTHFR[g]	1.57	1.21, 2.03	0.0006
DNA repair	Cao 2006	*65*	462	511	XRCC1	XRCC1[h]	0.48	0.27, 0.86	0.01*
	Cho 2003	*66*	334	283	XRCC1, hOGG1	XRCC1[i]	0.64	0.43, 0.96	not reported
	Qin 2011	*70*	755	755	88 DNA repair genes	RAD51L1[j,k]	1.22	1.04, 1.43	0.0017 in discovery stage[k]
	Yang 2007	*64*	153	168	XRCC1, XRCC3, XPD	XRCC1	1.83	1.29, 2.60	not reported
	Yang 2008	*68*	153	168	XPC	XPC[l]	1.60	1.16, 2,22	0.005
	Yang 2009	*67*	267	304	ERCC1	ERCC1[m]	1.41	1.08, 1.85	0.014
	Zheng 2007	*69*	531	480	N4BP2	None[n]	na	na	na
	Zheng 2011	*63*	1052	1168	NBS1	NBS1[o]	1.92 (het); 2.21 (homo)	1.33, 2.70; 1.48, 3.26	P_{trend}<0.0001
Tumor suppressor/ Oncogene	Duh 2004	*91*	55	114	FUS2	none	na	na	na
	Feng 2008	*92*	320	201	DCL-1	none	na	na	na
	Ren 2005	*93*	82	80	Tx	Tx[p]	not reported	not reported	0.007
Immunologic	Ben Nasr 2007	*120*	160	169	Il-8	Il-8	2.46	1.25, 4.88	0.004
	Farhat 2008	*117*	163	164	IL-18	none	na	na	na
	Gao 2009	*116*	206	373	IL-16	IL-16	1.67	1.18, 2.36	0.004
	Hassen 2007	*119*	206	155	TAP1	TAP1 (two loci)[q]	0.58; 0.52	0.38, 0.90; 0.33, 0.82	0.009, 0.002
	He 2007	*118*	434	512	TLR3	TLR3	1.49	1.10,2.00	0.0068
	Hirunsatit 2003	*124*	175	317	CR2, PIGR	PIGR	2.71	1.72-4.23	0.00001
	Ho 2006	*122*	89	360	TNFA	none	na	na	na

(continued)

Table 4.1 (continued)

	Study (first author and year)	Ref	Cases	Cont.	Genes studied	Significant gene associations	Odds ratio	95% CI	*P* value
	Jalbout 2003	*123*	140	274	TNFA, HSP70-2	HSP70-2	2.31	1.26, 4.22	0.006
	Nong 2009	*115*	250	270	IL-18	IL-18	1.70	1.66, 2.49	0.007
	Pratesi 2006	*121*	89	130	IL-10, IL-18	none	na	na	na
	Song 2006	*114*	486	529	TLR4	TLR4[r]	2.15	1.31, 3.51	0.01
	Sousa 2011	*107*	123	627	TNFA	TNFA	2.46	0.98, 6.17	0.047
	Tsai 2002a	*29*	47	119	TNFA	none	na	na	na
	Wei 2007a	*111*	280	290	Il-8	Il-8	1.40	1.06, 1.83	0.016
	Wei 2007b	*112*	189	210	IL-10	IL-10	2.25	1.53, 2.13	0.001
	Wei 2010	*108*	180	200	IL-2	IL-2	1.58	1.19, 2.13	0.002
	Xiao 2009	*109*	457	485	CTLA-4	CTLA-4	1.83	1.16,2.93	0.015
	Zhou 2006	*113*	487	580	TLR10	TLR10[s]	2.66 (for haplotype)	1.34, 3.82	0.0007
	Zhu 2008	*110*	113	144	IL-1B	IL-1B	1.53	1.07, 2.17	0.018
Genes in NPC-associated region	Guo 2006	*143*	350	288	PGM2, ARHH, APBB2, PHOX2B, KCTD8, GABRG1, USP46, SCFD2, CHIC2, GSH2	14 loci across region[t]	na	na	14 loci associated at *P* <0.05.[t]
	Li 2011	*127*	360	360	15 genes in 6p21.2-p23	GABBR1, HLA-A, HCG9	not reported	not reported	0.0004*, 0.0005*, 0.0017*

**P* value is corrected for multiple comparisons

[a] Possible interaction with smoking

[b] Negative interaction with smoking. No interactions with alcohol consumption

[c] Associations were significant only among smokers. No interacted with salted-fish or salted-vegetables. Findings were consistent with a parallel family-based study. False-positive report probability for six SNPs was <0.015

[d] No interaction with smoking. Possible interaction with alcohol

[e] Marginally significant associations were found only for specific strata of histological type and age

[f] Mutiple associated loci detected

[g] Interaction with smoking

[h] Possible interaction with smoking

[i] Nonsignificant main effect for hOGG1. An interaction of high risk alleles with CYP2E1 was reported

[j] Association was validated

[k] Possible interaction with salted fish and smoking. Validation had 1568 cases/1297 controls, and a Bonferroni corrected $P = 0.0381$

[l] No interaction found with either gender or smoking

[m] No interactions found with gender, smoking, or alcohol

[n] None of the SNPs were associated by themselves, but two haplotypes were differentially distributed between cases and controls

[o] This report shows variant to be functional in a cell transfection transcription assay

[p] Only chi-square analysis of genotype distributions between cases and controls was reported. No ORs reported

[q] TAP genes are located in HLA class II region

[r] Variate shown to be functional

[s] None of the SNPs were associated by themselves, but a haplotype was associated (adjusted $P=0.0007$)

[t] No loci retained significance after multiple comparison correction

for only four genes (*XRCC1*, *XPC*, *ERCC1*, and *RAD51L1*). One of these genes, *XRCC1*, was reported to be significantly associated in three different studies (Cho et al. 2003; Yang et al. 2007; Cao et al. 2006). And in one of those studies, *XRCC1* significance survived even after Bonferroni correction for multiple comparisons (Cao et al. 2006). However, the same variant allele (194Trp; rs1799782) was reported to be associated with risk (OR homozygous variant = 4.79; 95%CI = 1.48,15.52) in one study (Yang et al. 2007), while associated with protection (OR homozygous variant = 0.48; 95%CI = 0.27,0.86) in another (Cao et al. 2006). A third study (Cho et al. 2003) reported a protective association for a different allelic variant of *XRCC1* (280His, rs25489), but this failed to validate in one of the other two studies (Yang et al. 2007). And a recent study that genotyped 13 haplotype-tagging SNPs across the entire *XRCC1* gene (Qin et al. 2011) failed to detect any significant associations with NPC (see below). These differences in qualitative and quantitative association findings for *XRCC1*, some for the exact same alleles, raise doubts about biological relevance of these statistically significant associations. So despite the three separate reports of *XRCC1* variant alleles being associated with NPC, a role of *XRCC1* in NPC risk remains questionable.

In a recent investigation of 88 DNA repair genes, including *XRCC1*, *XPC*, and *ERCC1*, multiple haplotype-tagging SNPs were used to cover the entire sequence of each gene (Qin et al. 2011). Seven SNPs within three different genes (*RAD51L1*, *BRCA2*, *TP53BP1*) were found to be significantly associated with NPC in the discovery stage (cases/controls = 755/755). However, in the subsequent validation stage in a separate study population (cases/controls = 1,568/1,297), only two SNPs that were in strong LD with each other ($r^2 = 0.7$) maintained significance. These SNPs were both within the *RAD51L1* gene, which codes for a protein important for regulation of homologous recombinational DNA repair. Interestingly, a recent three-stage GWAS of breast cancer (cases/controls = 9,770/10,799) mapped the susceptibility locus to *RAD51L1* (Thomas et al. 2009b), supporting a very important role for this DNA repair gene in carcinogenesis. Conversely, the well-characterized homologous recombinational DNA repair and familial breast cancer risk gene, *BRCA2*, had two SNPs that associated with NPC in the discovery stage of this study; however, both failed to validate. Nevertheless, these similar genetic findings for the two cancers suggest a potential commonality in the etiology of NPC and breast cancer, at least in terms of DNA repair, and support the notion that dysfunctional homologous recombinational DNA repair promotes cancer risk.

4.3.4 Cytokines and Growth Factors

Various cytokines stimulate cell growth and proliferation and are thought to play important roles in the carcinogenic phenotype for several cancers, and cytokines are known to interact with EBV-infected cells (Mosialos 2001; Kis et al. 2006). *VEGF* and *EGF* have been reported to be modulated by EBV infection (Miller et al. 1995;

Tao et al. 2004; Stevenson et al. 2005; Krishna et al. 2006; Kung et al. 2011), and these have received some attention in NPC studies.

There were four studies of cytokines and growth factors (Wei et al. 2007c; Gao et al. 2008; Nasr et al. 2008; Wang et al. 2009a). Two studies reported significant associations between NPC and *VEGF* (Nasr et al. 2008; Wang et al. 2009a), but both had only marginal significance ($P<0.030$ and $P<0.029$), and neither was corrected for multiple comparisons, which would have extinguished their significant. A study of TNF-beta1 showed associations with NPC at two different loci with similar point estimates (1.63 and 1.70, respectively) and P values (0.009 and 0.006, respectively) (Wei et al. 2007c). But none of these studies have been validated.

4.3.5 Cell Adhesion

Proteins that play a role in cell adhesion often contribute to immunological function, stem cell differential, and tumor metastasis (Hirohashi and Kanai 2003; Crowson et al. 2007; Madson and Hansen 2007; Watt et al. 2008; Florian and Geiger 2010). Two association studies focused on the possible association of cell adhesion genes with NPC. In one study, the promoter region of the dendritic cell-specific intercellular adhesion molecule 3-grabbing non-integrin (DC-SIGN) – a pathogen recognition receptor that plays an important role in the susceptibility to various infectious diseases – was sequenced in 444 NPC patients and 464 controls (Xu et al. 2010). Results showed a highly significant protective haplotype (OR = 0.69; $P<0.0002$) that retained significant after 1,000 permutation test runs ($P<0.001$). This suggests that expression of the *DC-SIGN* gene may affect NPC susceptibility, possibly by modifying resistance to EBV infection.

In another study, the frequency of a variant of the E-cadherin gene promoter that had been demonstrated to modify gene expression during *in vitro* cell transfection assays (i.e., proved to be functional) was compared in 162 cases and 140 controls (Ben Nasr et al. 2010). Significantly increased risk of NPC for the variant carriers was observed (OR = 2.02; $P<0.008$). There was also a stronger association for NPC with the variant for early-onset (≤30 years old) NPC – OR = 3.86; $P<0.001$ – which is consistent with genetically based risk (Hemminki et al. 2004).

4.3.6 Tumor Suppressor Genes and Oncogenes

Tumor suppressor genes and oncogenes are carcinogenesis genes, and they are always prime candidates for cancer association studies. *TP53* is the most well-characterized tumor suppressor gene, and it plays well-described roles in both apoptosis and cell cycle arrest. For this reason, NPC association studies of *TP53* were reviewed in the apoptosis and cell cycle arrest section above. But apart from

TP53, three other studies investigated potential associations between carcinogenesis genes (FUS2, DCL-1, and Tx) and NPC (Duh et al. 2004; Ren et al. 2005; Feng et al. 2008). Of these genes, a significant association was only reported for a variant of the *Tx* gene ($P < 0.007$) (Ren et al. 2005). The *Tx* gene is a transforming gene that was isolated from an NPC cell line by DNA transfection and cloning techniques (Li et al. 2001). Bioinformatics approaches have shown the transforming gene to be an aberrant immunoglobulin kappa light chain gene containing a constant region, five intact joining regions, and five recombination signal sequences, but lacking the normal variable regions. The fact that this alternative *in vitro* screening approach has identified a gene with immunological function as a novel NPC tumor suppressor gene supports the notion that immune genes may affect NPC risk (see below). Nevertheless, the CGAS that reported the NPC risk association for *Tx* was quite small (82 cases/80 controls) and has not yet been validated.

4.3.7 DNA Methylation

A number of studies have suggested that epigenetic factors influence gene expression in NPC (Lo and Huang 2002; Fendri et al. 2009, 2010; Niller et al. 2009; Wang et al. 2009b, 2010a). Furthermore, promoter methylation is thought to be an important epigenetic mechanism for controlling gene expression in most cancers (Watanabe and Maekawa 2011), and EBV has been shown to interact with cellular DNA methylation processes (Niller et al. 2009). Nevertheless, only one study has looked at the DNA methylation pathways for candidate NPC genes (Cao et al. 2010a). That study revealed a highly significant association between an allele of the methylenetetrahydrofolate reductase (*MTHFR*) gene and NPC ($p < 0.0006$). There also was an indication of an interaction with smoking. *MTHFR* plays an important role in converting folate into a donor for DNA methylation, and thus could dysregulate DNA methylation patterns. However, these reported associations with NPC have not yet been validated.

4.3.8 Immunological Functions

HLA class I genes reside in a highly polymorphic gene region on chromosome 6 (6p21.3) and encode the proteins responsible for presenting foreign antigens to the immune system. As early as 1974, *HLA* variants were implicated in NPC risk (Simons et al. 1974), and in 1990 an *HLA*-linked loci was reported to be associated with a 21-fold increase in risk (Lu et al. 1990). Because of the connection between NPC and EBV infection, the notion of host immunological genes affecting NPC risk has been considered mechanistically plausible and etiologically attractive, and many studies have focused on *HLA* associations. But there have been some obstacles to their interpretation. Although, certain HLA class I alleles have been consistently shown to be associated with NPC risk, the reported associations are often race, ethnicity, or geographic region dependent. In addition, the *HLA* region has been disproportionately interrogated relative to the rest of the genome, suggesting that there might be elevated false-positive rates due to multiple comparisons, and likely some publication bias. Lastly, the *HLA* alleles associated with NPC are in LD with other genes, both immunological and nonimmunological, inside and outside the *HLA* region. For the reasons above, definitive conclusions about the role of HLA genes in NPC have been elusive.

There are two recent comprehensive reviews of the findings from *HLA* studies (Hassen et al. 2009; Li et al. 2009), so those studies are not reviewed here. But we address below whether the recent CGAS and GWAS support an association between immunological genes and NPC, and whether they inform our understanding of the role of immunologic genes in general, or HLA genes in particular, in NPC risk.

A total of 19 CGAS have looked at various immune pathway genes (Tsai et al. 2002a; Hirunsatit et al. 2003; Jalbout et al. 2003; Ho et al. 2006; Pratesi et al. 2006; Song et al. 2006; Zhou et al. 2006; Ben Nasr et al. 2007; Hassen et al. 2007; He et al. 2007; Wei et al. 2007a, 2007b, 2010b; Farhat et al. 2008; Zhu et al. 2008; Gao et al. 2009; Nong et al. 2009; Xiao et al. 2009; Sousa et al. 2010), and 15 different immune genes were studied. The interleukin genes were the largest group of immunological genes investigated. Nine studies looked at a total of six interleukin genes (*IL-1B*, *IL-2*, *IL-8*, *IL-10*, *IL-16*, *IL-18*), and all of the genes were reported to be associated with NPC in at least one study (Table 4.1). However, only one gene, IL-8, was reported to be associated with NPC in two separate studies (Ben Nasr et al. 2007; Wei et al. 2007b).

The Toll-like receptors (TLRs) were another group of immunological genes that received attention. TLRs play an essential role in initiating the immune response against pathogens and can recognize a wide variety of pathogen-associated molecular patterns from bacteria, viruses, and fungi (de la Barrera et al. 2006). For this reason, TLRs were considered candidate genes. To date, three different TLR genes (*TLR-3*, *TLR-4*, *TLR-10*) were investigated in three different studies (Song et al. 2006; Zhou et al. 2006; He et al. 2007), and all were reported to be associated with NPC. In contrast, the *TNFA* gene was investigated in four studies, but only one study found a significant association (Sousa et al. 2010), and even that association was marginal ($P < 0.047$).

The most highly significant association for an immunological gene was reported for the *PIGR* gene ($P < 0.00001$), which also had the largest reported effect size (OR = 2.71; 95%CI = 1.72, 4.23). The *PIGR* gene is part of the immunoglobulin superfamily and encodes a poly-Ig receptor that binds to polymeric immunoglobulin molecules at the basolateral surface of epithelial cells (Brandtzaeg 2009). Once bound, the complex is then transported across the cell to ultimately be secreted at the apical surface. *PIGR* has a role in maintaining mucosal immunity, including mucus tissues of the nasopharynx. So it is possible that *PIGR* can modify susceptibility to EBV infection, and this may support a role for HLA genes, although a direct connection between HLA genes and *PIGR* has not been established.

Another study took a somewhat different candidate-gene approach. These investigators interrogated 15 genes within the 6p21.3 chromosomal region, regardless of their putative function. They found highly associated SNPs in three genes from this region – *GABBR1*, *HLA-A*, and *HCG9* – with relatively low Bonferroni-corrected P values (0.0004, 0.0005, 0.0017, respectively). These findings strongly support the notion that the 6p21.3 region associates with NPC. But, because of high LD across the region, it is not clear whether these genes are in the NPC causal pathway or just represent good markers for a still unknown causative locus within the region.

In conclusion, the 19 CGAS that focused on immunological genes provide some supportive evidence for associations of immunological genes with NPC. However, with the possible exceptions of *IL-8* and *PIGR*, which had duplicate reports and a very low P value, respectively, the evidence is not very compelling. Few of the 19 studies corrected for multiple comparisons, nor did any validate their findings. And none investigated a possible interaction between the allegedly associated gene and EBV infection or exposure. Also, significant associations between NPC and genetic markers in genes selected because of their location within the 6p21.3 region further support the importance of this chromosomal region to NPC development, but do not inform us on the importance of their specific gene function to NPC. Taken together, these studies of immunological gene associations neither supported nor detracted from the proposition that HLA genes influence NPC risk.

4.4 Genome-Wide Association Studies

Two GWAS have focused on NPC endemic populations – one Taiwanese (Tse et al. 2009) and the other Southern Chinese (Bei et al. 2010). The Taiwanese study had 288 NPC cases and 297 controls, while the larger Cantonese study had 1,583 cases and 1,897 controls (in the discovery stage). Despite the differences in sample sizes, both studies identified their most significant signal in the *HLA* region (6p21) (Fig. 4.1).

The Taiwanese GWAS (Tse et al. 2009) were the first investigation to identify *GABBR1* at 6p21.31 as a promising candidate gene. Furthermore, the difference in the expression levels *GABBR1* between NPC tumors and the adjacent normal epithelial tissues suggested an importance of *GABBR1* in development of NPC. More interestingly, when the *GABBR1* transcript and protein levels in NPC cell lines were examined, downregulation of *GABBR1* protein in two NPC cell lines (AA genotype at rs29232) was observed compared with the immortalized nasopharyngeal epithelial cell line NP69 (AG genotype at rs29232). The risk allele of rs29232 was "A," and thus the homozygous carrier of A allele exhibited a lower protein level than the heterozygous carrier. On the other hand, the Taiwanese study did not compare the *GABBR1* transcript and protein expression levels between normal and cancer cell lines. Therefore, more work is needed to elucidate the relationship between the carriers of the "A" allele and levels of gene expression. In a follow-up study carried out by another group (Li et al. 2011), there was shown to be a

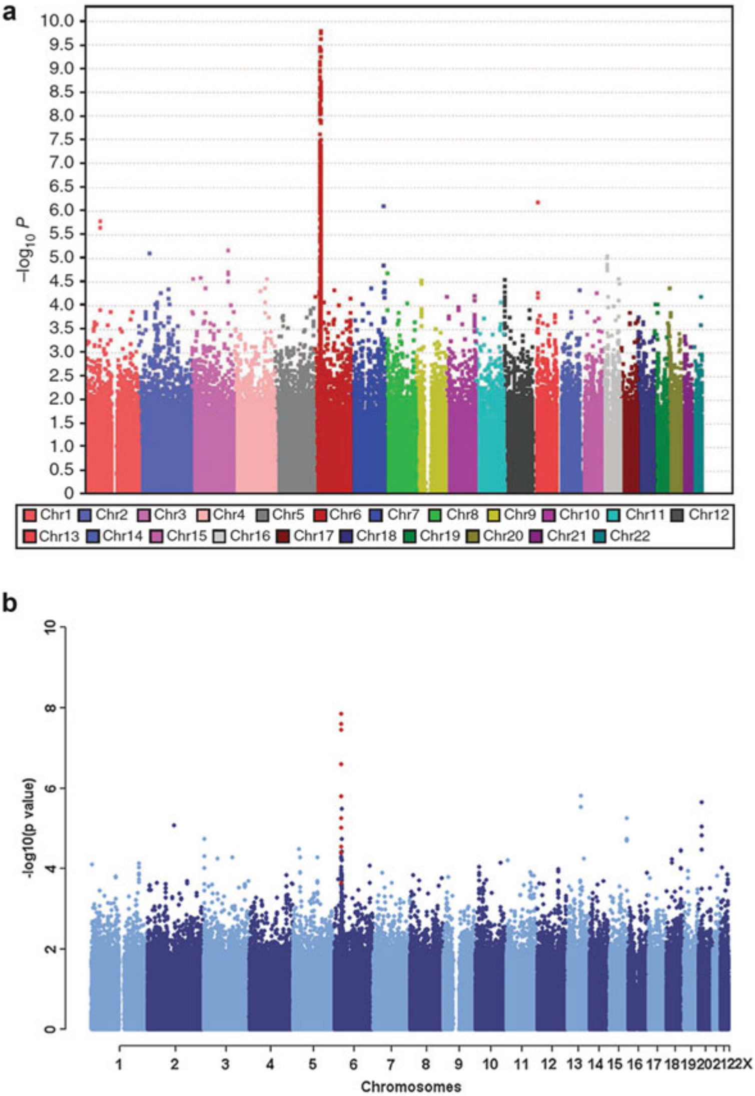

Fig. 4.1 GWAS showing evidence for the association of *HLA* and nasopharyngeal carcinoma risk. *Panel a*. Manhattan plot of the genome-wide P values of association for the mainland GWA study in Southern China (Bei et al. 2010). *Panel b*. Manhattan plot of the genome-wide P values of association for the GWA study in Taiwan (Tse et al. 2009)

downregulation of *GABBR1* transcripts in NPC tumors, which may suggest that downregulation of *GABBR1* expression is one of the tumorigenic mechanisms. However, *GABBR1* encodes a G-protein-coupled subunit of the gamma-aminobutyric acid (GABA) B receptor 1. Its ligand – gamma-aminobutyric acid (GABA) – is the main inhibitory neurotransmitter in central nervous system and is not known to have a role in nonneuronal tissue. So it is difficult to envisage how the *GABBR1* gene might affect NPC risk. Nevertheless, in tissue expression comparisons, T and B lymphocytes have the next highest levels of *GABBR1* expression after neuronal tissues (Burren et al. 2010; T1DBase team 2011), suggesting a role for GABBR1 in immune function. Regardless of its mechanism, *GABBR1*'s possible involvement in NPC etiology warrants further research.

The Cantonese study (Bei et al. 2010) also found an association within the *HLA* region on 6p21. Further, they reported three novel NPC susceptibility loci on 3q26, 9p21, and 13q12 and identified several novel risk genes: *TNFRSF19* (tumor necrosis factor receptor superfamily, member 19), *MDS1-EVI1* (a zinc-finger DNA-binding transcription activator), and the *CDKN2A-CDKN2B* gene cluster (cyclin-dependent kinases involved in cell cycle arrest). All of these genes have previously been shown to be involved with leukemia, supporting their role in carcinogenesis. And it has been shown that NPC patients are at higher risk of developing leukemia (Scelo et al. 2007), so it can be hypothesized that NPC and leukemia may share common genetic risk factors. But it is possible as well that EBV infection is a risk factor for both NPC and leukemia (Tedeschi et al. 2007). It is also notable that the *CDKN2A-CDKN2B* gene cluster is deleted in about 40% of NPC tumors, suggesting a potential tumor suppressor function at this locus (Lo and Huang 2002).

4.5 Discussion

There have been multiple CGAS of NPC that have used pathway-based approaches to select candidate genes for interrogation, and a number of SNP variants have been reported to be statistically significantly associated with NPC. Most of these associations have had small effect sizes and marginal statistical significance, which might be expected based on what we already know about SNP associations with disease and the statistical power needed to detect those associations (Park et al. 2010). Nevertheless, it is the prevalence of these variants in the population, rather than the magnitude of the effect sizes, which drives their potential relevance to the attributable risk of NPC. Of more concern is the fact that few of the reported gene variant associations have been validated in a second study population, and very few have been shown to be biologically functional or in LD with any functional locus, leaving most reported associations unconfirmed and inconclusive.

Earlier family studies have linked *HLA* loci with NPC risk, and this has precipitated a large number of CGAS that have focused on genes involved in immunological functions both inside and outside of the *HLA* region. Although these studies have reported some associations between immunological gene variants and NPC, they cannot be considered independent confirmations of immunologically based risk, because the immunological genes have been disproportionately interrogated relative to the rest of the genome, so there is an oversampling bias for immune genes. Also, there has not been any obvious patterns of association for the immunological genes, and the reported variants are often synonymous coding variants, or in introns or other noncoding sequences. This suggests that they must be in LD with an unidentified functional variant in neighboring sequences, if they are truly associated with NPC risk.

In contrast, GWAS have independently confirmed an association between the *HLA* region and NPC risk, but have shed no further insight on mechanisms. The associated GWAS markers are unlikely to directly impact function themselves, again suggesting that they are in LD with yet to be identified functional loci. None of the genes with CGAS reports of associations have turned up in the GWAS, and the genes that have been found to be associated with NPC by GWAS were not interrogated in any of the CGAS reports. Thus, there have been no cross confirmations between CGAS and GWAS. Failure for GWAS to confirm associations reported by CGAS does not invalidate the CGAS findings, since a variety of factors can influence the sensitivity of GWAS to detect any particular associated SNP. Thus far, GWAS investigations of various diseases have only been successful in confirming those candidate-gene associations that had very large effect sizes (Siontis et al. 2010). A contributing factor to the paucity of confirmations by GWAS is that CGAS, unlike GWAS, do not use standardized platforms and procedures, making direct comparisons between GWAS and CGAS difficult. Nevertheless, the lack of confirmation with GWAS is disheartening.

As for the NPC-associated genes identified by GWAS, only a couple seem to be involved in the major candidate pathways, and it is not immediately obvious how their known or proposed functions directly modify NPC risk. Thus, they do not appear to inform our current understanding of the carcinogenic mechanisms of NPC. Again, the gene function associated with the genetic marker needs to be identified and characterized in order to capitalize on the discovered association with NPC, even if the association findings are valid.

Regarding GWAS confirmation of NPC's association with the *HLA* locus, this finding is gratifying but anticipated. The previous association with *HLA* found through family studies is so strong and reproducible (Li et al. 2009) that it is hard to see how this strong association would not be seen with GWAS. But the GWAS findings do not provide us with any higher resolution of the disease region than the linkage analyses do. So GWAS do not bring us any closer than before to the risk gene in the *HLA* region. Also, it is not even clear that risk associated with the *HLA* region has anything to do with HLA genes. This region of the genome is rich in genes and rich in diseases that associate with it, including multiple sclerosis, epilepsy, schizophrenia, Hodgkin and non-Hodgkin lymphomas, chronic lymphocytic leukemia, and breast cancer (McKnight et al. 2009; Hawthorn et al. 2010; Meng et al. 2010; Slager et al. 2010; Vrzalova et al. 2010; Wang et al. 2010b; Zollino et al. 2010; McElroy and Oksenberg 2011; Moutsianas et al. 2011), and most of these diseases are not thought to be primarily due to an *HLA* dysfunction. With the advent in whole-genome sequencing technology, we anticipate that there will be better definitions for NPC-relevant haplotypes in the *HLA* region, and further biological mechanism related to NPC will be clarified with the emergence of reliable haplotypes and adequate sample sizes in future studies.

It may be that our knowledge of NPC disease etiology is too imperfect to reliably identify likely biochemical pathways for risk modification. In the advent of GWAS of NPC, perhaps the best use of the candidate-gene approach is to perform high-density SNP interrogations within genomic regions of interest as identified by GWAS. This is the approach taken in two NPC association reports (Guo et al. 2006; Li et al. 2011). One of these (Li et al. 2011) was able to confirm in a different population the association of NPC with the *GABBR1* gene that was discovered in an earlier GWAS (Tse et al. 2009). This association withstood even Bonferroni correction for multiple comparisons and is, therefore, quite robust despite the modest effect size (OR = 1.67; 95%CI = 1.48, 1.88) of the original GWAS report.

Another value of validation by CGAS is that it can typically be achieved in a different and smaller study population. These smaller populations are much more likely to have complete and useful environmental exposure data, which in turn provides the potential to assess possible gene-environment interactions. Although it should also be possible to explore gene-environmental interactions with GWAS, there is seldom adequate exposure information for these larger, often pooled, study populations. Environmental exposure data allow for adjustments for the environmental risk factors and also for assessment of gene-environment interactions in a way that is typically not achievable in large GWAS. Controlling for environmental risk factors may have the added advantage in that it may boost the power to detect the genetic associations. This would be particularly relevant for NPC, where multiple environmental risk factors are known and there are geographic pockets of populations at risk. Nevertheless, few of the CGAS to date have utilized environmental data in their analytical design. Doing so could significantly augment the value the CGAS approach for NPC.

Clearly, GWAS have provided an avenue for evaluation of the association between common genetic variants and human diseases. However, most variants identified by GWAS seem to be merely markers rather than being causal for disease, and this is undoubtedly the case for NPC. We also know that for the diseases with large heritability estimates (i.e., 60–80%) such as NPC, only 5–10% of that heritability has been found by GWAS.

The main limitations of GWAS are the following: (1) *Low power due to the issue of multiple testing.* To increase the power, populations with large sample sizes might help to solve these problems. Although the cost of genotyping has been reduced dramatically with the advances of technology, collecting large numbers of patients will still be an obstacle. In addition, the power of an interaction study in for GWAS dataset is typically low, and analyzing large number of variables in various combinations becomes computationally challenging. (2) *Population differences.* Some SNPs that are tightly associated with a disease in one population may be only weakly associated with the same disease in other populations. Since many GWAS are based on case–control designs, the effect of population admixture could be substantial, and the association, to a large extent, may depend on ethnicity-related factors. (3) *GWAS are mainly focused on single-nucleotide variations.* Copy number variations (CNVs), structural variations (SVs), and deletions have received less attention, and (4) gene-gene interactions and gene-environment interactions have often been neglected. In most GWAS, due to the small effect sizes of common SNPs, methods used for detecting potential interactions are typically underpowered. Large sample sizes and improved analytical techniques might ease these problems.

The limitations of GWAS compel epidemiologists and geneticists to further consider the contributions of CNVs, SVs (Bansal et al. 2010), gene-gene and gene-environment interactions, and, in particular, the joint contribution of rare variants (frequency less than 1%) to human diseases (Bansal et al. 2010). The advent of revolutionary high-throughput sequencing technology (also called "next-generation" sequencing or NGS, paralleled sequencing) has paved a way for a better understanding of the origins of human cancer. As a superior model to study *HLA* and virus infection and environment-virus-gene interaction, it is plausible to conduct genetic study on NPC using next-generation sequencing. The interpretation of carcinogenesis of NPC might largely depend on acquiring genetic information from both virus and the host, and also the elucidation of their interactions with environmental risk factors.

Finally, causal variants for NPC will only be found by complete genomic sequencing of cases and controls. Currently, we still need to rely on the CGAS and GWAS to identify smaller genomic regions where we can focus our sequencing efforts. To achieve this goal, CGAS, GWAS, and NGS need to be harmonized with each other in order the extract the most information possible from the limited number of populations available for study. In this regard, the power limitations of GWAS due to multiple-comparison corrections should be taken into account, and some consideration should be afforded even to nonsignificant multiple-comparison-adjusted SNPs if their effects sizes are large or if the findings are supportive of an earlier reported CGAS association. Likewise, CGAS should incorporate the current GWAS platform markers, in order to validate reported GWAS associations. If this is not possible, then analyzing highly correlated SNPs may still allow informative cross comparisons between CGAS, GWAS, and NGS results.

In short, GWAS should not be viewed as superseding CGAS in the search for NPC-associated genetic variants, since both approaches have their strengths and weaknesses. However, it is relatively easier to replicate findings in independent GWAS than in CGAS. CGAS findings are often harder to be replicated due to the difference in platforms, imperfect tagging in some of the studies, and impact of population stratification. (In CGAS, researchers do not typically have a large enough number of SNPs to correct for potential population stratification.) Still, the two approaches should be viewed as complementary to each other and preliminary to direct sequencing. In the advent of GWAS technology, the best use of CGAS may be to confirm GWAS findings by blanketing the region of interest with high-density SNP coverage, and thereby validating the GWAS association, while also setting the stage for subsequent validation by deep sequencing.

The biggest challenge ahead for NPC is likely to be the characterization of gene-environmental interactions. In light of the very high prevalence of EBV infection within the high-risk populations, it may be difficult to achieve the power necessary to demonstrate interactions between EBV and genetic factors, unless the interactions are very strong. Unfortunately, the potential strength of interactions is something that cannot either be assessed or predicted, based on current data from either CGAS or GWAS, and statistical methodologies for quantifying and assessing interactions have not yet been validated. Given the presumed necessity that persons at genetic risk of NPC avoid environmental NPC exposure risks, the importance of this information to targeting public health prevention interventions cannot be overstated and is an area that warrants further scientific attention.

Acknowledgements Author Affiliations: The Fisher Center for Familial Cancer Research, Lombardi Cancer Center, Georgetown University Medical Center, Washington, DC 20007, USA (Timothy J. Jorgensen); Unit of Statistical Genomics, Intramural Research Program, National Institute of Mental Health, National Institute of Health, Bethesda, MD 20852, USA (Yin Yao Shugart). Cancer Epidemiology Program, Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore MD 21205, USA (Timothy J. Jorgensen).

The views expressed in this presentation do not necessarily represent the views of the NIMH, NIH, HHS, or the United States Government.

Conflicts of Interest: None declared.

References

Bailey SG, Verrall E, et al. Functional interaction between Epstein-Barr virus replication protein Zta and host DNA damage response protein 53BP1. J Virol. 2009;83(21):11116–22.

Bansal V, Libiger O, et al. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85.

Bei JX, Li Y, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet. 2010;42(7):599–603.

Ben Nasr H, Chahed K, et al. Association of IL-8 (−251)T/A polymorphism with susceptibility to and aggressiveness of nasopharyngeal carcinoma. Hum Immunol. 2007;68(9):761–9.

Ben Nasr H, Hamrita B, et al. A single nucleotide polymorphism in the E-cadherin gene promoter −160 C/A is associated with risk of nasopharyngeal cancer. Clin Chim Acta. 2010;411(17–18): 1253–7.

Berwick M, Vineis P. Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. J Natl Cancer Inst. 2000;92(11):874–97.

Brandtzaeg P. Mucosal immunity: induction, dissemination, and effector functions. Scand J Immunol. 2009;70(6):505–15.

Brennan B. Nasopharyngeal carcinoma. Orphanet J Rare Dis. 2006;1:23.

Burren OS, Adlem EC, et al. T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. Nucleic Acids Res. 2010;39(Database issue):D997–1001.

Cao Y, Miao XP, et al. Polymorphisms of XRCC1 genes and risk of nasopharyngeal carcinoma in the Cantonese population. BMC Cancer. 2006;6:167.

Cao Y, Miao XP, et al. Polymorphisms of methylenetetrahydrofolate reductase are associated with a high risk of nasopharyngeal carcinoma in a smoking population from Southern China. Mol Carcinog. 2010a;49(11):928–34.

Cao Y, Miao XP, et al. Polymorphisms of death pathway genes FAS and FASL and risk of nasopharyngeal carcinoma. Mol Carcinog. 2010b;49(11):944–50.

Cao SM, Simons MJ, et al. The prevalence and prevention of nasopharyngeal carcinoma in China. Chin J Cancer. 2011;30(2):114–19.

Chang ET, Adami HO. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 2006;15(10):1765–77.

Cheng YJ, Chien YC, et al. No association between genetic polymorphisms of CYP1A1, GSTM1, GSTT1, GSTP1, NAT2, and nasopharyngeal carcinoma in Taiwan. Cancer Epidemiol Biomarkers Prev. 2003;12(2):179–80.

Cheung HW, Chun AC, et al. Inactivation of human MAD2B in nasopharyngeal carcinoma cells leads to chemosensitization to DNA-damaging agents. Cancer Res. 2006;66(8):4357–67.

Chew MM, Gan SY, et al. Interleukins, laminin and Epstein – Barr virus latent membrane protein 1 (EBV LMP1) promote metastatic phenotype in nasopharyngeal carcinoma. BMC Cancer. 2010;10:574.

Cho EY, Hildesheim A, et al. Nasopharyngeal carcinoma and genetic polymorphisms of DNA repair enzymes XRCC1 and hOGG1. Cancer Epidemiol Biomarkers Prev. 2003;12(10): 1100–4.

Chou J, Lin YC, et al. Nasopharyngeal carcinoma – review of the molecular mechanisms of tumorigenesis. Head Neck. 2008;30(7):946–63.

Crowson AN, Magro C, et al. The molecular basis of melanomagenesis and the metastatic phenotype. Semin Oncol. 2007;34(6):476–90.

de la Barrera S, Aleman M, et al. Toll-like receptors in human infectious diseases. Curr Pharm Des. 2006;12(32):4173–84.

Deng L, Zhao XR, et al. Cyclin D1 polymorphism and the susceptibility to NPC using DHPLC. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai). 2002;34(1):16–20.

Dodd LE, Sengupta S, et al. Genes involved in DNA repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 2006;15(11):2216–25.

Duh FM, Fivash M, et al. Characterization of a new SNP c767A/T (Arg222Trp) in the candidate TSG FUS2 on human chromosome 3p21.3: prevalence in Asian populations and analysis of association with nasopharyngeal cancer. Mol Cell Probes. 2004;18(1):39–44.

Farhat K, Hassen E, et al. Functional IL-18 promoter gene polymorphisms in Tunisian nasopharyngeal carcinoma patients. Cytokine. 2008;43(2):132–7.

Fendri A, Masmoudi A, et al. Inactivation of RASSF1A, RARbeta2 and DAP-kinase by promoter methylation correlates with lymph node metastasis in nasopharyngeal carcinoma. Cancer Biol Ther. 2009;8(5):444–51.

Fendri A, Khabir A, et al. Epigenetic alteration of the Wnt inhibitory factor-1 promoter is common and occurs in advanced stage of Tunisian nasopharyngeal carcinoma. Cancer Invest. 2010;28(9):896–903.

Feng XL, Zhou W, et al. The DLC-1–29A/T polymorphism is not associated with nasopharyngeal carcinoma risk in Chinese population. Genet Test. 2008;12(3):345–9.

Florian MC, Geiger H. Concise review: polarity in stem cells, disease, and aging. Stem Cells. 2010;28(9):1623–9.

Frenkel K. Carcinogen-mediated oxidant formation and oxidative DNA damage. Pharmacol Ther. 1992;53(1):127–66.

Gajwani BW, Devereaux JM, et al. Familial clustering of nasopharyngeal carcinoma. Cancer. 1980;46(10):2325–7.

Gallicchio L, Matanoski G, et al. Adulthood consumption of preserved and nonpreserved vegetables and the risk of nasopharyngeal carcinoma: a systematic review. Int J Cancer. 2006;119(5): 1125–35.

Gao LB, Wei YS, et al. No association between epidermal growth factor and epidermal growth factor receptor polymorphisms and nasopharyngeal carcinoma. Cancer Genet Cytogenet. 2008;185(2):69–73.

Gao LB, Liang WB, et al. Genetic polymorphism of interleukin-16 and risk of nasopharyngeal carcinoma. Clin Chim Acta. 2009;409(1–2):132–5.

Grimm T, Schneider S, et al. EBV latent membrane protein-1 protects B cells from apoptosis by inhibition of BAX. Blood. 2005;105(8):3263–9.

Gruhne B, Sompallae R, et al. Three Epstein-Barr virus latency proteins independently promote genomic instability by inducing DNA damage, inhibiting DNA repair and inactivating cell cycle checkpoints. Oncogene. 2009;28(45):3997–4008.

Guo XC, Scott K, et al. Genetic factors leading to chronic Epstein-Barr virus infection and nasopharyngeal carcinoma in South East China: study design, methods and feasibility. Hum Genomics. 2006;2(6):365–75.

Guo X, O'Brien SJ, et al. GSTM1 and GSTT1 gene deletions and the risk for nasopharyngeal carcinoma in Han Chinese. Cancer Epidemiol Biomarkers Prev. 2008;17(7):1760–3.

Guo X, Zeng Y, et al. Genetic Polymorphisms of CYP2E1, GSTP1, NQO1 and MPO and the risk of nasopharyngeal carcinoma in a Han Chinese population of Southern China. BMC Res Notes. 2010;3:212.

Hadhri-Guiga B, Toumi N, et al. Proline homozygosity in codon 72 of TP53 is a factor of susceptibility to nasopharyngeal carcinoma in Tunisia. Cancer Genet Cytogenet. 2007;178(2):89–93.

Haorah J, Zhou L, et al. Determination of total N-nitroso compounds and their precursors in frankfurters, fresh meat, dried salted fish, sauces, tobacco, and tobacco smoke particulates. J Agric Food Chem. 2001;49(12):6068–78.

Hassen E, Farhat K, et al. TAP1 gene polymorphisms and nasopharyngeal carcinoma risk in a Tunisian population. Cancer Genet Cytogenet. 2007;175(1):41–6.

Hassen E, Nahla G, et al. The human leukocyte antigen class I genes in nasopharyngeal carcinoma risk. Mol Biol Rep. 2009;37(1):119–26.

Hatzivassiliou E, Mosialos G. Cellular signaling pathways engaged by the Epstein-Barr virus transforming protein LMP1. Front Biosci. 2002;7:d319–29.

Hawthorn L, Luce J, et al. Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. BMC Cancer. 2010;10:460.

He JF, Jia WH, et al. Genetic polymorphisms of TLR3 are associated with Nasopharyngeal carcinoma risk in Cantonese population. BMC Cancer. 2007;7:194.

He Y, Zhou GQ, et al. Correlation of polymorphism of the coding region of glutathione S- transferase M1 to susceptibility of nasopharyngeal carcinoma in South China population. Ai Zheng. 2009;28(1):5–7.

Hemminki K, Rawal R, et al. Genetic epidemiology of cancer: from families to heritable genes. Int J Cancer. 2004;111(6):944–50.

Hildesheim A, Chen CJ, et al. Cytochrome P4502E1 genetic polymorphisms and risk of nasopharyngeal carcinoma: results from a case–control study conducted in Taiwan. Cancer Epidemiol Biomarkers Prev. 1995;4(6):607–10.

Hildesheim A, Anderson LM, et al. CYP2E1 genetic polymorphisms and risk of nasopharyngeal carcinoma in Taiwan. J Natl Cancer Inst. 1997;89(16):1207–12.

Hildesheim A, Dosemeci M, et al. Occupational exposure to wood, formaldehyde, and solvents and risk of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 2001;10(11):1145–53.

Hirohashi S, Kanai Y. Cell adhesion system and human cancer morphogenesis. Cancer Sci. 2003;94(7):575–81.

Hirunsatit R, Kongruttanachok N, et al. Polymeric immunoglobulin receptor polymorphisms and risk of nasopharyngeal cancer. BMC Genet. 2003;4:3.

Ho SY, Wang YJ, et al. Evaluation of the associations between the single nucleotide polymorphisms of the promoter region of the tumor necrosis factor-alpha gene and nasopharyngeal carcinoma. J Chin Med Assoc. 2006;69(8):351–7.

Hou DF, Wang SL, et al. Expression of CYP2E1 in human nasopharynx and its metabolic effect in vitro. Mol Cell Biochem. 2007;298(1–2):93–100.

Iwakawa M, Goto M, et al. DNA repair capacity measured by high throughput alkaline comet assays in EBV-transformed cell lines and peripheral blood cells from cancer patients and healthy volunteers. Mutat Res. 2005;588(1):1–6.

Jalbout M, Bouaouina N, et al. Polymorphism of the stress protein HSP70-2 gene is associated with the susceptibility to the nasopharyngeal carcinoma. Cancer Lett. 2003;193(1):75–81.

Jeyakumar A, Brickman TM, et al. Review of nasopharyngeal carcinoma. Ear Nose Throat J. 2006;85(3):168–70, 172–3, 184.

Ji MF, Wang DK, et al. Sustained elevation of Epstein-Barr virus antibody levels preceding clinical onset of nasopharyngeal carcinoma. Br J Cancer. 2007;96(4):623–30.

Jia WH, Collins A, et al. Complex segregation analysis of nasopharyngeal carcinoma in Guangdong, China: evidence for a multifactorial mode of inheritance (complex segregation analysis of NPC in China). Eur J Hum Genet. 2005;13(2):248–52.

Jia WH, Pan QH, et al. A case–control and a family-based association study revealing an association between CYP2E1 polymorphisms and nasopharyngeal carcinoma risk in Cantonese. Carcinogenesis. 2009;30(12):2031–6.

Jiang JH, Jia WH, et al. Genetic polymorphisms of CYP2A13 and its relationship to nasopharyngeal carcinoma in the Cantonese population. J Transl Med. 2004;2(1):24.

Jorgensen TJ, Ruczinski I, et al. Hypothesis-driven candidate gene association studies: practical design and analytical considerations. Am J Epidemiol. 2009;170(8):986–93.

Kis LL, Takahara M, et al. Cytokine mediated induction of the major Epstein-Barr virus (EBV)-encoded transforming protein, LMP-1. Immunol Lett. 2006;104(1–2):83–8.

Krishna SM, James S, et al. Expression of VEGF as prognosticator in primary nasopharyngeal cancer and its relation to EBV status. Virus Res. 2006;115(1):85–90.

Kung CP, Meckes Jr DG, et al. Epstein-Barr virus LMP1 activates EGFR, STAT3, and ERK through effects on PKCdelta. J Virol. 2011;85(9):4399–408.

Li M, Ren W, et al. Nucleotide sequence analysis of a transforming gene isolated from nasopharyngeal carcinoma cell line CNE2: an aberrant human immunoglobulin kappa light chain which lacks variable region. DNA Seq. 2001;12(5–6):331–5.

Li X, Fasano R, et al. HLA associations with nasopharyngeal carcinoma. Curr Mol Med. 2009;9(6):751–65.

Li S, Deng Y, et al. Diagnostic value of Epstein-Barr virus capsid antigen-IgA in nasopharyngeal carcinoma: a meta-analysis. Chin Med J (Engl). 2010;123(9):1201–5.

Li Y, Fu L, et al. Identification of genes with allelic imbalance on 6p associated with nasopharyngeal carcinoma in southern Chinese. PLoS One. 2011;6(1):e14562.

Liu MT, Chen YR, et al. Epstein-Barr virus latent membrane protein 1 induces micronucleus formation, represses DNA repair and enhances sensitivity to DNA-damaging agents in human epithelial cells. Oncogene. 2004;23(14):2531–9.

Liu MT, Chang YT, et al. Epstein-Barr virus latent membrane protein 1 represses p53-mediated DNA repair and transcriptional activity. Oncogene. 2005;24(16):2635–46.

Liu JP, Cassar L, et al. Mechanisms of cell immortalization mediated by EB viral activation of telomerase in nasopharyngeal carcinoma. Cell Res. 2006;16(10):809–17.

Lo KW, Huang DP. Genetic and epigenetic changes in nasopharyngeal carcinoma. Semin Cancer Biol. 2002;12(6):451–62.

Lu SJ, Day NE, et al. Linkage of a nasopharyngeal carcinoma susceptibility locus to the HLA region. Nature. 1990;346(6283):470–1.

Madson JG, Hansen LA. Multiple mechanisms of Erbb2 action after ultraviolet irradiation of the skin. Mol Carcinog. 2007;46(8):624–8.

Martin D, Gutkind JS. Human tumor-associated viruses and new insights into the molecular mechanisms of cancer. Oncogene. 2008;27 Suppl 2:S31–42.

McElroy JP, Oksenberg JR. Multiple sclerosis genetics 2010. Neurol Clin. 2011;29(2):219–31.

McKnight AJ, Currie D, et al. Targeted genome-wide investigation identifies novel SNPs associated with diabetic nephropathy. Hugo J. 2009;3(1–4):77–82.

Meng H, Powers NR, et al. A dyslexia-associated variant in DCDC2 changes gene expression. Behav Genet. 2010;41(1):58–66.

Miller WE, Earp HS, et al. The Epstein-Barr virus latent membrane protein 1 induces expression of the epidermal growth factor receptor. J Virol. 1995;69(7):4390–8.

Mosialos G. Cytokine signaling and Epstein-Barr virus-mediated cell transformation. Cytokine Growth Factor Rev. 2001;12(2–3):259–70.

Moutsianas L, Enciso-Mora V, et al. Multiple Hodgkin lymphoma-associated loci within the HLA region at chromosome 6p21.3. Blood. 2011;118(3):670–4.

Nasr HB, Chahed K, et al. Functional vascular endothelial growth factor −2578 C/A polymorphism in relation to nasopharyngeal carcinoma risk and tumor progression. Clin Chim Acta. 2008;395(1–2):124–9.

Nazar-Stewart V, Vaughan TL, et al. Glutathione S-transferase M1 and susceptibility to nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev. 1999;8(6):547–51.

Ng CC, Yew PY, et al. A genome-wide association study identifies ITGA9 conferring risk of nasopharyngeal carcinoma. J Hum Genet. 2009;54(7):392–7.

Niller HH, Wolf H, et al. Epigenetic dysregulation of the host cell genome in Epstein-Barr virus-associated neoplasia. Semin Cancer Biol. 2009;19(3):158–64.

Nong LG, Luo B, et al. Interleukin-18 gene promoter polymorphism and the risk of nasopharyngeal carcinoma in a Chinese population. DNA Cell Biol. 2009;28(10):507–13.

O'Nions J, Allday MJ. Epstein-Barr virus can inhibit genotoxin-induced G1 arrest downstream of p53 by preventing the inactivation of CDK2. Oncogene. 2003;22(46):7181–91.

Pang MF, Lin KW, et al. The signaling pathways of Epstein-Barr virus-encoded latent membrane protein 2A (LMP2A) in latency and cancer. Cell Mol Biol Lett. 2009;14(2):222–47.

Park JH, Wacholder S, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010;42(7):570–5.

Parkin DM, Iscovich J. Risk of cancer in migrants and their descendants in Israel: II. Carcinomas and germ-cell tumours. Int J Cancer. 1997;70(6):654–60.

Pokrovskaja K, Okan I, et al. Epstein-Barr virus infection and mitogen stimulation of normal B cells induces wild-type p53 without subsequent growth arrest or apoptosis. J Gen Virol. 1999;80(Pt 4):987–95.

Pratesi C, Bortolin MT, et al. Interleukin-10 and interleukin-18 promoter polymorphisms in an Italian cohort of patients with undifferentiated carcinoma of nasopharyngeal type. Cancer Immunol Immunother. 2006;55(1):23–30.

Qin HD, Shugart YY, et al. Comprehensive pathway-based association study of DNA repair gene variants and the risk of nasopharyngeal carcinoma. Cancer Res. 2011;71(8):3000–8.

Ren W, Zheng H, et al. A functional single nucleotide polymorphism site detected in nasopharyngeal carcinoma-associated transforming gene Tx. Cancer Genet Cytogenet. 2005;157(1):49–52.

Rodriguez-Antona C, Gomez A, et al. Molecular genetics and epigenetics of the cytochrome P450 gene family and its relevance for cancer risk and treatment. Hum Genet. 2009;127(1):1–17.

Rowe DT. Epstein-Barr virus immortalization and latency. Front Biosci. 1999;4:D346–71.

Saha A, Murakami M, et al. Epstein-Barr virus nuclear antigen 3C augments Mdm2-mediated p53 ubiquitination and degradation by deubiquitinating Mdm2. J Virol. 2009;83(9):4652–69.

Scelo G, Boffetta P, et al. Second primary cancers in patients with nasopharyngeal carcinoma: a pooled analysis of 13 cancer registries. Cancer Causes Control. 2007;18(3):269–78.

Sckolnick J, Murphy J, et al. Microsatellite instability in nasopharyngeal and lymphoepithelial carcinomas of the head and neck. Am J Surg Pathol. 2006;30(10):1250–3.

Shen GP, Pan QH, et al. Human genetic variants of homologous recombination repair genes first found to be associated with Epstein-Barr virus antibody titers in healthy Cantonese. Int J Cancer. 2011;129(6):1459–66.

Simons MJ. Nasopharyngeal carcinoma as a paradigm of cancer genetics. Chin J Cancer. 2011;30(2):79–84.

Simons MJ, Day NE, et al. Nasopharyngeal carcinoma V: immunogenetic studies of Southeast Asian ethnic groups with high and low risk for the tumor. Cancer Res. 1974;34(5):1192–5.

Siontis KC, Patsopoulos NA, et al. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. Eur J Hum Genet. 2010;18(7):832–7.

Slager SL, Rabe KG, et al. Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. Blood. 2010;117(6):1911–16.

Song C, Chen LZ, et al. Functional variant in the 3′-untranslated region of Toll-like receptor 4 is associated with nasopharyngeal carcinoma risk. Cancer Biol Ther. 2006;5(10):1285–91.

Sousa H, Santos AM, et al. Linkage of TP53 codon 72 pro/pro genotype as predictive factor for nasopharyngeal carcinoma development. Eur J Cancer Prev. 2006;15(4):362–6.

Sousa H, Breda E, et al. Genetic risk markers for nasopharyngeal carcinoma in Portugal: tumor necrosis factor alpha -308G >A polymorphism. DNA Cell Biol. 2010;30(2):99–103.

Stevenson D, Charalambous C, et al. Epstein-Barr virus latent membrane protein 1 (CAO) up-regulates VEGF and TGF alpha concomitant with hyperlasia, with subsequent up-regulation of p16 and MMP9. Cancer Res. 2005;65(19):8826–35.

T1DBase team. 2011. T1D/Base: GABBR1. Retrieved January 3, 2012, URL: http://t1dbase.org/page/Overview/display/gene_id/54393

Tam JS, Murray HG. Nasopharyngeal carcinoma and Epstein-Barr virus–associated serologic markers. Ear Nose Throat J. 1990;69(4):261–7.

Tao Y, Song X, et al. Nuclear translocation of EGF receptor regulated by Epstein-Barr virus encoded latent membrane protein 1. Sci China C Life Sci. 2004;47(3):258–67.

Tedeschi R, Bloigu A, et al. Activation of maternal Epstein-Barr virus infection and risk of acute leukemia in the offspring. Am J Epidemiol. 2007;165(2):134–7.

Thomas DC, Conti DV, et al. Use of pathway information in molecular epidemiology. Hum Genomics. 2009a;4(1):21–42.

Thomas G, Jacobs KB, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 2009b;41(5):579–84.

Tiwawech D, Srivatanakul P, et al. The p53 codon 72 polymorphism in Thai nasopharyngeal carcinoma. Cancer Lett. 2003;198(1):69–75.

Tiwawech D, Srivatanakul P, et al. Glutathione S-transferase M1 gene polymorphism in Thai nasopharyngeal carcinoma. Asian Pac J Cancer Prev. 2005;6(3):270–5.

Tiwawech D, Srivatanakul P, et al. Cytochrome P450 2A6 polymorphism in nasopharyngeal carcinoma. Cancer Lett. 2006;241(1):135–41.

Tsai MH, Chen WC, et al. Correlation of p21 gene codon 31 polymorphism and TNF-alpha gene polymorphism with nasopharyngeal carcinoma. J Clin Lab Anal. 2002a;16(3):146–50.

Tsai MH, Lin CD, et al. Prognostic significance of the proline form of p53 codon 72 polymorphism in nasopharyngeal carcinoma. Laryngoscope. 2002b;112(1):116–19.

Tse KP, Su WH, et al. Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am J Hum Genet. 2009;85(2):194–203.

Vrzalova Z, Hruba Z, et al. Chimeric CYP21A1P/CYP21A2 genes identified in Czech patients with congenital adrenal hyperplasia. Eur J Med Genet. 2010;54(2):112–17.

Wang T, Hu K, et al. Polymorphism of VEGF-2578C/A associated with the risk and aggressiveness of nasopharyngeal carcinoma in a Chinese population. Mol Biol Rep. 2009a;37(1):59–65.

Wang T, Liu H, et al. Methylation associated inactivation of RASSF1A and its synergistic effect with activated K-Ras in nasopharyngeal carcinoma. J Exp Clin Cancer Res. 2009b;28:160.

Wang S, Xiao X, et al. TFPI-2 is a putative tumor suppressor gene frequently inactivated by promoter hypermethylation in nasopharyngeal carcinoma. BMC Cancer. 2010a;10:617.

Wang SS, Menashe I, et al. Variations in chromosomes 9 and 6p21.3 with risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers Prev. 2010b;20(1):42–9.

Watanabe Y, Maekawa M. Methylation of DNA in cancer. Adv Clin Chem. 2011;52:145–67.

Watt FM, Estrach S, et al. Epidermal Notch signalling: differentiation, cancer and adhesion. Curr Opin Cell Biol. 2008;20(2):171–9.

Wei YS, Kuang XH, et al. Interleukin-10 gene promoter polymorphisms and the risk of nasopharyngeal carcinoma. Tissue Antigens. 2007a;70(1):12–7.

Wei YS, Lan Y, et al. Single nucleotide polymorphism and haplotype association of the interleukin-8 gene with nasopharyngeal carcinoma. Clin Immunol. 2007b;125(3):309–17.

Wei YS, Zhu YH, et al. Association of transforming growth factor-beta1 gene polymorphisms with genetic susceptibility to nasopharyngeal carcinoma. Clin Chim Acta. 2007c;380(1–2):165–9.

Wei KR, Yu YL, et al. Epidemiological trends of nasopharyngeal carcinoma in China. Asian Pac J Cancer Prev. 2010a;11(1):29–32.

Wei YS, Lan Y, et al. Association of the interleukin-2 polymorphisms with interleukin-2 serum levels and risk of nasopharyngeal carcinoma. DNA Cell Biol. 2010b;29(7):363–8.

Wu CC, Liu MT, et al. Epstein-Barr virus DNase (BGLF5) induces genomic instability in human epithelial cells. Nucleic Acids Res. 2009;38(6):1932–49.

Xiao M, Qi F, et al. Functional polymorphism of cytotoxic T-lymphocyte antigen 4 and nasopharyngeal carcinoma susceptibility in a Chinese population. Int J Immunogenet. 2009;37(1):27–32.

Xiao M, Zhang L, et al. Genetic polymorphisms of MDM2 and TP53 genes are associated with risk of nasopharyngeal carcinoma in a Chinese population. BMC Cancer. 2010;10:147.

Xiong A, Clarke-Katzenberg RH, et al. Epstein-Barr virus latent membrane protein 1 activates nuclear factor-kappa B in human endothelial cells and inhibits apoptosis. Transplantation. 2004;78(1):41–9.

Xu YF, Liu WL, et al. Sequencing of DC-SIGN promoter indicates an association between promoter variation and risk of nasopharyngeal carcinoma in cantonese. BMC Med Genet. 2010;11:161.

Yang CS, Yoo JS, et al. Cytochrome P450IIE1: roles in nitrosamine metabolism and mechanisms of regulation. Drug Metab Rev. 1990;22(2–3):147–59.

Yang ZH, Du B, et al. Genetic polymorphisms of the DNA repair gene and risk of nasopharyngeal carcinoma. DNA Cell Biol. 2007;26(7):491–6.

Yang ZH, Liang WB, et al. The xeroderma pigmentosum group C gene polymorphisms and genetic susceptibility of nasopharyngeal carcinoma. Acta Oncol. 2008;47(3):379–84.

Yang ZH, Dai Q, et al. Association of ERCC1 polymorphisms and susceptibility to nasopharyngeal carcinoma. Mol Carcinog. 2009;48(3):196–201.

Yi F, Saha A, et al. Epstein-Barr virus nuclear antigen 3C targets p53 and modulates its transcriptional and apoptotic activities. Virology. 2009;388(2):236–47.

Yung WC, Ng MH, et al. p53 codon 72 polymorphism in nasopharyngeal carcinoma. Cancer Genet Cytogenet. 1997;93(2):181–2.

Zeng YX, Jia WH. Familial nasopharyngeal carcinoma. Semin Cancer Biol. 2002;12(6):443–50.

Zheng H, Li LL, et al. Role of Epstein-Barr virus encoded latent membrane protein 1 in the carcinogenesis of nasopharyngeal carcinoma. Cell Mol Immunol. 2007a;4(3):185–96.

Zheng MZ, Qin HD, et al. Haplotype of gene Nedd4 binding protein 2 associated with sporadic nasopharyngeal carcinoma in the Southern Chinese population. J Transl Med. 2007b;5:36.

Zheng J, Zhang C, et al. Functional NBS1 polymorphism is associated with occurrence and advanced disease status of nasopharyngeal carcinoma. Mol Carcinog. 2011;50(9):689–96.

Zhou XX, Jia WH, et al. Sequence variants in toll-like receptor 10 are associated with nasopharyngeal carcinoma risk. Cancer Epidemiol Biomarkers Prev. 2006;15(5):862–6.

Zhu Y, Xu Y, et al. Association of IL-1B gene polymorphisms with nasopharyngeal carcinoma in a Chinese population. Clin Oncol (R Coll Radiol). 2008;20(3):207–11.

Zhuo X, Cai L, et al. GSTM1 and GSTT1 polymorphisms and nasopharyngeal cancer risk: an evidence-based meta-analysis. J Exp Clin Cancer Res. 2009a;28:46.

Zhuo XL, Cai L, et al. TP53 codon 72 polymorphism contributes to nasopharyngeal cancer susceptibility: a meta-analysis. Arch Med Res. 2009b;40(4):299–305.

Zollino M, Gurrieri F, et al. Integrated analysis of clinical signs and literature data for the diagnosis and therapy of a previously undescribed 6p21.3 deletion syndrome. Eur J Hum Genet. 2010;19(2):239–42.

Chapter 5
QTL Mapping of Molecular Traits for Studies of Human Complex Diseases

Chunyu Liu

Abstract Genetic mapping of quantitative trait loci (QTL) offers a powerful and efficient approach to discover putative regulatory regions of traits and to define novel functional implications of genetic variants. Here we reviewed recent progress on QTL mapping of molecular traits, including gene expression, DNA methylation, as well as protein expression, metabolites. QTL mapping of molecular traits has better chance to succeed in relatively small sample size study as fewer nongenetic factors or gene-gene interactions may involve. Knowledge derived from QTL mapping will help us to uncover understanding of biology in complex traits and diseases and enhance power of genetic association study. In the context of study of complex diseases, we focused on expression QTL and methylation QTL, presenting major findings and technique considerations, including experimental platform, sample quality, size, and heterogeneity, as well as analytical procedure and significance criteria. Lastly, we discussed the current and future use of QTL data in study of complex diseases.

Keywords Complex diseases • DNA methylation • eQTL • mQTL • pQTL

5.1 Introduction

Complex diseases, such as diabetes, Crohn's disease, asthma, and many neuropsychiatric diseases, have multiple genetic and environmental factors involved. Although genetic contribution is apparent, transmission in families do not obey the Mendelian rules of inheritance. High prevalence in population, strong heterogeneity,

C. Liu (✉)
Department of Psychiatry, University of Illinois at Chicago,
900 Ashland Ave. Room 1006, Chicago, IL 60607, USA
e-mail: liucy@uic.edu

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_5, © Springer Science+Business Media Dordrecht 2012

incomplete penetrance, and complex spectrum of phenotypes are frequently observed for these diseases. Identification of their genetic factors promises to bring us better understanding of the disease etiology, new treatment, and most importantly personalized medicine. But the path reaching this goal is not easy. Actually, it is much more difficult than study of rare Mendelian disorders. Prior to 2005, genome-wide linkage and association studies were thought to be the silver bullets to nail down all the common risk genes. Unfortunately, the reality showed us the complexity beyond what we have expected.

5.1.1 Genome-Wide Association Study and Its Limitation

Genome-wide linkage and association studies made full use of the gradually improved genetic map of human genome. With millions of genetic variants, particularly single nucleotide polymorphisms (SNPs) identified throughout human genome, Affymetrix and Illumina provide affordable SNP microarray or BeadChip for "unbiased," hypothesis-free, genome-wide association test for study of any common diseases or traits.

Since 2005, with thousands even tens of thousands of samples recruited in each study, genome-wide association studies (GWAS) have made significant progress. NHGRI Catalog of Genome-Wide Association Studies (http://www.genome.gov/gwastudies) has collected more than 1,100 GWASs of more than 590 diseases or traits by the end of 2011. Except for a few diseases like age-related macular degeneration (ARMD, (Klein et al. 2005)), most of the diseases only have weak-effect loci revealed with odds ratio less than 2. "Missing heritability" has been the most complaint heard about GWAS (Manolio et al. 2009; Eichler et al. 2010). Actually, "missing biology" may be more problematic: Most of the discovered associations linking to SNPs do not have obvious biological functions as they are frequently located in intronic or noncoding regions. One example is the GWAS signal identified for the bipolar disorder as summarized in Table 5.1. Most of the associated SNPs are in intronic or intergenic regions with no obvious function.

Meanwhile, with the linkage disequilibrium (LD), a SNP association frequently cannot really pinpoint to a specific gene in a genomic region. Only until we have one specific gene and its causal functional variants actual being identified, we will be able to put together the puzzle pieces of the disease biology. The disease gene and biological pathway can then be revealed and followed-up.

One example is the synonymous coding variant in *PBRM1* gene, rs2251219, which was reported to be associated with bipolar and major depression by McMahon et al. (2010). It was replicated in bipolar but not in major depression (Breen et al. 2011). rs2251219 has a nearby nonsynonymous (V355M) variant, rs2289247, in the gene *GNL3* (GTPase nucleostemin), which was involved in proliferation of stem cells, especially in the central nervous system. Our analysis showed that rs2251219

Table 5.1 Bipolar disorder GWAS signals reaching genome-wide significance

Study	Gene	SNPs	Locations
PGC (Sklar et al. 2011)	CACNA1C, ODZ4	rs4765913; rs12576775	Intronic
Cichon et al. (2011)	NCAN	rs1064395	3′UTR
McMahon et al. (Baum et al. 2008)	PBRM1	rs2251219	Cds-synon
Wang et al. (2010)	ASTN2, GABRR1	rs11789399	Intergenic
Huang et al. (2010a)	ADM	rs6484218	Intergenic
Liu et al. (2011)	CACNA1C	rs1006737	Intronic
Ferreira et al. (2008)	ANK3	rs10994336	Intergenic
Baum et al. (2008)	DGKH	rs1012053	Intronic

is associated with expression of *GNL3* at both exon and transcript level, in both cerebellum and parietal cortex. Therefore, we propose that *GNL3* may be the actual risk bipolar disorder gene rather than *PBRM1*, although rs2251219 is 142 Kb away from rs2289247. This example also shows that an eQTL could be located right inside another gene. Different genes may share not only exons but also regulatory elements. Current SNP annotation using only physical location could be functionally misleading.

While researchers are still working hard to collect more samples to improve statistical power of GWAS, aiming to identify more weak-effect risk genes, integrating knowledge of biological functions of genetic variants into GWAS might be an important alternative approach to enhance GWAS power so that weak-effect risk genes can be discovered without increasing sample size.

Study of biological function of genetic variants will benefit both recovering missing biological mechanism and discovering of novel weak-effect risk genes.

5.1.2 Functionality of Genetic Variants

Genetic variants could have their functions defined at various biological levels, from molecular functions such as gene expression, protein and lipid level, cellular functions such as cell structure and nerve excitability, to tissue and organ functions such as brain activity, till high-order functions such as human cognitive and emotion behaviors. In general, the higher level the function is, the more genetic and environmental factors can be involved. Although some high-level functions could be products of relatively simple genetic variants, majority of the high-level functions such as human behaviors will have many genetic and environmental factors interplayed, consequently, have weaker correlations with genetic variants than gene expression measures do. It is natural to assume that many higher level functions are built upon organization of lower level functions. Therefore, study of biological functions at molecular level, which are in scope of many -omics, such as genomics and epigenomics, deemed to be more fruitful as bigger effect size of genetic variants is expected for those traits. These studies will also be essential for understanding of higher level phenotypes.

Here, we will focus on reviewing recent studies of SNP functions measured by genomic and epigenomic methods. Genetic mapping is making more and more contributions to the study of these functionalities, as it can discover novel functions of genetic variants more efficiently than traditional biochemical or mutagenesis, transgenic animal experiments. Certainly, similar to all other association tests, genetic mapping reveals the statistical correlation between measure of a quantitative trait and a genomic region. The correlation can only suggest but never prove a causal relationship. The actual biology, cause-consequence relationship has to be established through follow-up experiments.

5.1.3 QTL Mapping and Genetic Variants

A quantitative trait can be recorded as a continuous variable in a population. The earliest study of a quantitative trait was enzyme activity (Schwartz 1962). Genetically mapping quantitative traits, or quantitative trait loci (QTL), began in the 1980s since DNA markers were introduced.

Mapping of QTLs, just like other genetic traits, can use both linkage and association methods. Linkage includes variance components analysis, regression, and nonparametric methods. Association test can be either family-based test or population-based test. In general, successful association studies produce better resolution than successful linkage studies. This chapter focuses on GWAS mapping of QTL in human. QTL mapping can be performed in animal or other model species, like mouse or yeast. They are not covered here.

A very fruitful practice of QTL mapping so far is the mapping of gene expression quantitative traits loci (eQTLs). eQTL mapping started about 15 years ago (Damerval et al. 1994). After GWAS was implemented, eQTL mapping study bloomed. Other molecular QTL Including gene methylation QTL (mQTL), protein QTL (pQTL), and others, gradually have also been presented. Creative use QTL mapping is opening a broad venue toward understanding of biology and complex traits.

5.2 QTL Mapping of Molecular Traits

Molecular traits are defined as phenotypes that can be assessed, mostly quantitatively, at molecular level in contrast to morphological phenotypes and behavioral, psychological measures. Molecular traits include most of the molecules that are currently measured by biochemical and molecular biological methods, such as gene expression, DNA methylations, histone modifications, enzyme activity, hormones, and metabolites. Most of them are the causes also the products of gene-environment interaction at different levels (Fig. 5.1).

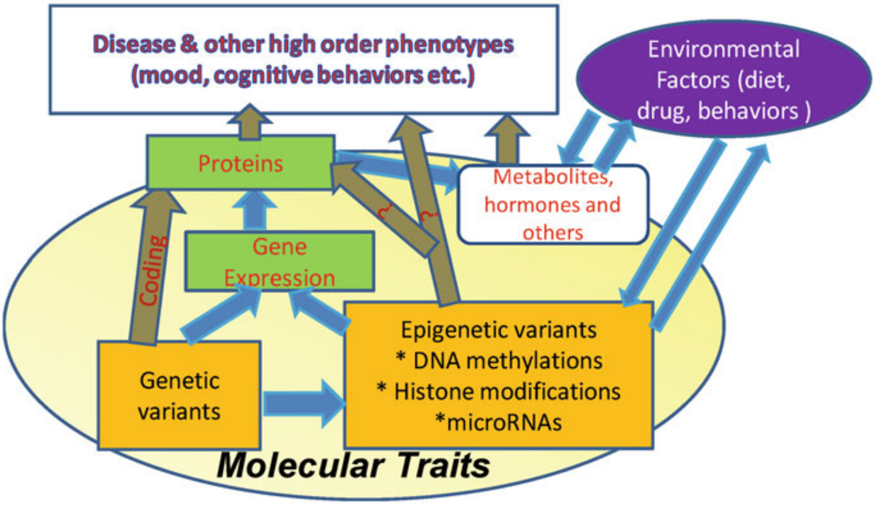
Disease & other high order phenotypes
(mood, cognitive behaviors etc.)
Environmental Factors (diet, drug, behaviors)
Proteins
Metabolites, hormones and others
Coding
Gene Expression
~
~
Epigenetic variants
* DNA methylations
* Histone modifications
*microRNAs
Genetic variants
Molecular Traits

Fig. 5.1 Molecular traits link DNA/RNA to environment and high-order phenotypes

5.2.1 eQTL

An eQTL refers to a genetic variant in which the genotypes are associated with differential gene expression. Through an eQTL mapping study, we can identify potential regulatory regions in the genome for expression of a specific gene. The simplest model is that the genetic variant is either located in a regulatory element or in LD with a variant in the element so that the DNA sequence change could affect transcription or degradation efficiency. And one needs to bear in mind that the actual causal relationship or regulatory machinery will not be apparent without additional experiments.

With millions of SNPs genotyped, a genome-wide eQTL study is normally performed by partitioning the tests into cis- and trans-tests (Fig. 5.2). cis- (or local) association is between expression level of one gene and a nearby SNP, one located within an arbitrarily defined distance such as 500 Kb or 1–2 Mb. Trans- (or distal) associations include all non-cis-pairs. The trans- can be associations between the expression of a gene on one chromosome and a SNP located on another chromosome.

HapMap lymphoblastoid cell lines (LCLs) have been the mostly studied samples for eQTL mapping (Monks et al. 2004; Morley et al. 2004; Stranger et al. 2005, 2007; Cheung et al. 2005; ; Storey et al. 2007; Veyrieras et al. 2008; Zhang et al. 2008). The other human tissues that have been studied include liver (Schadt et al. 2008), kidney (Wheeler et al. 2009), blood and subcutaneous adipose tissue (Emilsson et al. 2008), and whole blood (Fehrmann et al. 2011), and brain (Myers et al. 2007; Heinzen et al. 2008; Webster et al. 2009; Liu et al. 2010), omental adipose, subcutaneous adipose, and liver (Dobrin et al. 2011). LCLs from asthma patients (Dixon et al. 2007; Moffatt et al. 2007) and from twins (Min et al. 2011) have also been studied for eQTL.

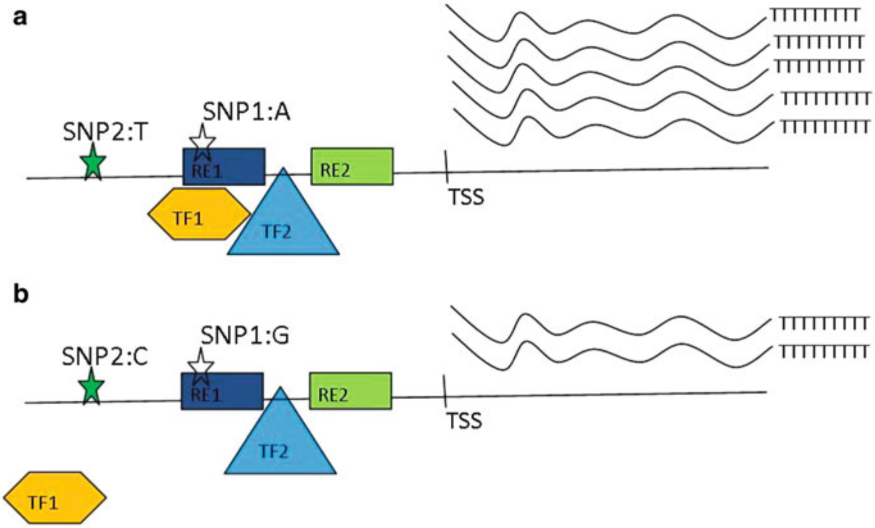

Fig. 5.2 Model of SNPs presenting eQTL in cis-association. SNP1 is located inside regulatory element1 (*RE1*). Its A allele has strong binding affinity to transcription factor1 (*TF1*). Its G allele does not bind TF1 well and consequently leading to reduced expression. SNP2 is in linkage disequilibrium with SNP1. Therefore, genotypes of both SNP1 and SNP2 show correlation with expression. *TSS* transcription start site

Several review articles have summarized part of the past eQTL studies (Cheung and Spielman 2009; Cookson et al. 2009; Liu 2011). An updated list of brain eQTL studies is shown in Table 5.2.

Brain has been the most intensively studied tissue next to the HapMap LCLs. These two tissues represent two extreme of eQTL mapping in terms of complexity. LCL has relatively uniform cell type, with many environmental influences washed out during culture. Brain tissue block could contain hundreds or more different cell types and may be affected by lifetime and postmortem environmental influences. Different brain regions are structurally and functionally different. LCL sample can be prepared freshly and easily. Human brain is rarely accessible alive. Because of the complexity and very limited access of brain, eQTL mapping in human brain is at its early stage.

Most of the published eQTL mapping studies were limited to the summarized measure of transcripts, averaging all the splicing forms of each gene. But It is estimated that 42–73% of human genes are alternatively spliced (Modrek et al. 2001; Johnson et al. 2003; Clark et al. 2007). Human brain carries even more tissue-specific alternative splice forms than other tissue (Xu et al. 2002; Johnson et al. 2009).

The heritability of splicing isoforms was first investigated in CEPH LCLs (Kwan et al. 2007) (Nembaware et al. 2008). Splicing eQTLs were also studied in CEPH LCLs using RNA-Seq (Pickrell et al. 2010; Montgomery et al. 2010). Hundreds of eQTLs for quantification of exon or whole gene transcripts were identified by these two studies. There are more eQTLs for exons detected than for whole transcripts.

Table 5.2 eQTL studies on human postmortem brains

Authors	Samples	Platforms	Findings with definition of significance
Myers et al. (2007)	193 neuropathologically normal human brain, frontal, temporal, and parietal regions (European descent)	Affymetrix 500K, Illumina HumanRefseq-8 Expression array	Significant cis-associations for 433 SNP-transcript pairs, 336 SNP-transcript pairs show trans-association (transcript-specific empirical *P* value ≤ 0.05). 25 SNP-transcript pairs, involving 2 genes, KIF1B and IPP, show significant cis-association after further correction for the number of phenotypes tested
Webster et al. (2009)	363 cortical samples from brains Alzheimer patients (European descent)	Affymetrix 500K, Illumina HumanRefseq-8 expression array	The expression levels of 9% cortical transcripts had expression profiles correlated with cis-SNP genotypes at a region-wide or genome-wide significance level
Heinzen et al. (2008)	93 normal frontal cortical brains and 80 normal mononucleated blood cell samples (no ethnic info)	Affymetrix Human Exon 1.0 ST array and Illumina Human Hap550K chips	23 "high confidence associations" with total transcript expression and 80 associations with specific exons. Fewer than 50% of the implicated SNPs show effects in both brain and blood
Liu et al. (2010)	127 prefrontal cortex samples, including bipolar, schizophrenia, depression, and controls from the Stanley collections	Affymetrix 5.0 array and U133A array	The cis-analysis revealed 562 associations involving 106 genes that remained significant after correcting for all the expression phenotypes and all the SNPs tested for each gene. In the trans-analysis, 241 associations involving 157 genes reached a genome-wide significance level, but none survived additional correction for the number of expression phenotypes tested
Gibbs et al. (2010)	Four human brain regions each: cerebellum, frontal cortex, temporal cortex, and pons regions from 150 individuals (600 samples total) (European descent)	Infinium HumanHap550 Beadchips; HumanRef-8 Expression BeadChips	2,944–4,781 cis-associations and 471–1,826 trans-associations are significant (permutation correction for SNPs tested, FDR for traits tested)
Kang et al. (2011)	57 normal individuals, with 15-period system spanning the periods from embryonic development to late adulthood, of 16 brain regions; mixed ethnicities	Illumina 2.5-million SNP chip; Affymetrix GeneChip Human Exon 1.0 ST array	2–39 cis-eQTLs in different brain regions. (Bonferroni correction for SNP tested; FDR $q < 0.1$ for genome-wide)

(continued)

Table 5.2 (continued)

Authors	Samples	Platforms	Findings with definition of significance
Colantuoni et al. (2011)	269 individuals, prefrontal cortex, (Primary African American and Caucasian samples)	Either Illumina Infinium II 650K or Illumina Infinium HD Gemini 1M Duo BeadChips for genotype; Custom expression array	1,628 individual associations surpass Bonferroni correction for all the SNPs and traits tested ($p<2.6e{-}12$).
Liu et al. (unpublished)	146 parietal cortex samples and 131 cerebellum samples from SMRI (European descent)	Affymetrix 5.0 array for genotyping; Human Gene 1.0 ST array for expression	6,794 significant cis-eQTLs, 991 significant trans-eQTLs in parietal cortex region and 9,010 significant cis-eQTLs, 960 significant trans-eQTLs in cerebellum region (phenotype-wide p value<0.05)

Note: All these studies used microarrays probing largely the same 20–30 thousand human genes, but only Human Gene 1.0 or Exon 1.0 ST array provide information of individual exons. But exon-level analysis was not summarized here. The numbers of SNPs tested vary by genotyping platforms

Many factors determine the number of eQTLs that can be discovered. They include (1) experimental platform, (2) RNA quality, (3) sample heterogeneity, (4) sample size, (5) covariates, (6) data analytical procedures, and (7) significance criteria.

5.2.1.1 Experimental Platforms

There are three technologies measuring mRNA expression, including microarray or BeadChip, real-time quantitative PCR (qPCR), and RNA-Seq.

Illumina, Affymetrix, and Agilent are the major vendors of microarray technology. Although they all designed array to probe the 30,000 human genes, different microarray designs have pros and cons for their use of different numbers of probes on each transcript or exon, for the different lengths of oligo probes, and for their signal detection methods. Expense is certainly another important factor in the option of platform. One major selling point of Affymetrix Human Gene or Exon 1.0 ST array is that they provide decent coverage of individual known exons so that expression of specific splicing isoforms can be assayed and evaluated for eQTL mapping.

All microarray technology share built-in limitations due to being hybridization based. The oligonucleotide probes may hybridize to duplicated or repeat sequence or hit genomic regions with SNPs in populations (Alberts et al. 2007; Duan et al. 2008; Gamazon et al. 2010), which will affect hybridization efficiency. In turn, false positives and false negatives can be produced. Ideally, all the probes containing SNPs should be excluded from analysis. We established a database for the list of expression microarray probes containing common SNPs at http://bioinfo.psych.uic.edu/ArrayGenes/SNPsInProbes.jsp. Additionally, detection of the fluorescence signals has a limited dynamic range so that the measure will not be accurate at the low or high ends or out of, the linear correlation (dynamic) range. Lastly, microarray can only measure the expression of known targets. Novel transcripts and exons will be the blind spot to microarray.

The qPCR method has a wider dynamic range but may still be affected by SNPs in primers or in TaqMan probe-binding sites. Therefore, the results have potential to be false because of a poor primer or probe design. Like microarray, qPCR is also limited to known targets.

With a high price tag, RNA-Seq has significant advantages over traditional expression microarrays and the qPCR method. The dynamic range of RNA-Seq is reported to be at least 8,000-fold, a vast improvement over the 60-fold of DNA microarrays (Nagalakshmi et al. 2008). Montgomery SB et al. have found that approximately ten million reads of sequence can provide a comparable dynamic range as a microarray (Montgomery et al. 2010). RNA-Seq's measure of expression will not be affected by SNPs. Instead, allelic expression can be directly measured as sequence variants detected in the transcripts (Heap et al. 2010). Most uniquely, RNA-Seq allows the identification of novel transcripts and splicing isoforms. Several investigations (Marioni et al. 2008; Wang et al. 2009) have demonstrated the feasibility of using RNA-Seq to profile gene expression in eQTL mapping. The first two RNA-Seq-based eQTL studies used HapMap LCLs (Pickrell et al. 2010;

Montgomery et al. 2010) and identified over 100 novel putative protein-coding exons and over 1,000 genes with eQTLs at gene or splice variant expression levels. Majewski J. and Pastinen T. had a thorough review of RNA-Seq application in eQTL mapping (Majewski and Pastinen 2011). As the costs of next-generation sequencing gradually decrease, RNA-Seq is expected be used more in eQTL mapping studies.

5.2.1.2 RNA Quality

RNA quality is critical for eQTL as it affects accuracy of measurement of expression. RNA degrades rapidly, and tissues need to be quickly collected and processed carefully. For this reason, studies utilizing tissues collected from living body or cultured cells should produce higher quality data in general than using postmortem tissues. RNA integrity number (RIN) is a frequently used index of RNA quality (Schroeder et al. 2006).

5.2.1.3 Sample Heterogeneity

Sample heterogeneity involves several levels. One tissue may contain many different cell types. Different tissues or cell types could have different gene expression profiles. Many eQTLs are thus tissue or cell type specific. Study showed that LCL and whole blood have distinct eQTL profile (Powell et al. 2011). As discussed above, brain is a particularly complex tissue while thousands of cell types blended in the "soup." Some tissues such as leukocyte could be more accessible and homogeneous.

In the Multiple Tissue Human Expression Resource (MuTHER) study, three tissues (156 LCL, 160 skin and166 fat) from the same individuals of healthy female twins were used for *cis*-eQTL analysis. This study demonstrates that 30% of eQTLs are shared among tissues, while 29% are exclusively tissue-specific. Even for shared eQTLs, 10–20% have significant tissue differences (Nica et al. 2011).

Genetic heterogeneity is another layer of complexity investigators have to deal with in eQTL mapping. Population structure, difference of minor allele frequency in different ethnic populations, could affect eQTL mapping like all other GWASs. Hsiao et al. carefully evaluated the effects in their study (Hsiao et al. 2010).

Mixing heterogeneous samples into one study could lead to increased power to detecting shared eQTLs after careful controlling the population structure issue, but it will overestimate power for detecting population-unique eQTLs.

5.2.1.4 Sample Size

Sample size is an obvious determining factor for statistic power in eQTL mapping. The more samples used, the more eQTLs can be detected, assuming the other factors are fixed. Based on published studies, one needs less than 100 samples to identify those very strong cis-eQTLs. When thousands of samples are recruited for eQTL mapping, we can expect that most of the transcripts in human genome will reveal their eQTLs.

Trans-eQTLs requires larger sample collection. With 1,469 unrelated blood samples, high-quality trans-associations were detected and replicated in a different set of tissues and sample collections (Fehrmann et al. 2011).

5.2.1.5 Covariates

Covariates may impact on association tests. Lab experiments are subject to batch effects, which are systematic, nonbiological variations among experimental batches. Since eQTL mapping requires relatively large sample size, measures of expression data of all samples in one batch is practically infeasible. In order to minimize batch effects, universal technical replicates could be used in all batches to evaluate batch effects. Each batch should contain both cases and controls for analysis involving case–control comparison to minimize the confounding bias. A number of algorithms are available for removing potential batch effects from expression data, and our systematic evaluation (Chen et al. 2011a) has found ComBat (Johnson et al. 2007) to be the best.

Sample demographic information, clinical measures are also important covariates as they may influence gene expression. In study of brain eQTL, postmortem interval (PMI) and brain pH are important covariates. Study of cultured cell line may have some advantages as many environmental factors, covariates, could be washed off during the culture. Study of 47 monozygotic twin pairs did not detect significant contribution of 14 blood biochemical traits and cell count on gene expression in whole blood and LCL culture (Powell et al. 2011). The covariates should be evaluated carefully before putting them aside.

5.2.1.6 Analytical Procedures

In data analysis, quality control is the first important thing to do for obtaining reliable results. Having discussed above, removing probes that might be affected by common SNPs, or nonspecific binding from the analysis, controlling batch effects and covariates are important. Surrogate Variable Analysis (SVA) (Leek and Storey 2007) is a good software to regress out both known and unknown covariates so that the residues can be used for eQTL mapping as two studies have used (Liu et al. 2010; Colantuoni et al. 2011). It could be considered to be a method to obtain robust eQTL mapping in samples confounded with other covariates, like affection status, and brain pH. New method has been developed and to be test in actual eQTL study (Listgarten et al. 2010).

Since genotypic data is used in the study, population stratification should also be considered in the association tests in a more serious manner when heterogeneous population is used.

5.2.1.7 Significance Criteria

Significance criteria are important for reducing false calling of eQTLs. Because of simultaneous tests of large amount of associations, multiple testing may lead to false positives without proper correction. Bonferroni correction, permutation, or false discovery rate (FDR) have been commonly used. In our own study, we defined two levels of significance: region-wide or genome-wide significance referring to the adjusted *p* for controlling for all the SNPs tested for cis- or trans-association tests, respectively. Phenotype-wide significance refers to the adjusted *p* after additional control for the number of expression traits studied.

It is worth mentioning that the significance in replicate study could be relaxed depending on the number of positive findings in the initial discovery studies. The direction of association is also very important. Findings that can be replicated in multiple datasets will be more credible.

5.2.2 mQTL

DNA methylation is an important epigenetic modification on DNA nucleotides without changing the actual sequence. It normally occurs at the CpG site changing Cytosine to 5-methylcytosine (5mC). DNA methylation is classically considered as a major gene expression regulator: Higher methylation represses gene expression. This simple relationship is gradually being broken down by the recent findings after the research studies extended into non-promoter regions (Jones 1999; Deng et al. 2009; Ball et al. 2009; Rauch et al. 2009). Studies showed that highly expressed genes tend to have extensive gene-body methylation and minimal promoter methylation, whereas the bodies of weakly expressed genes are less methylated (Deng et al. 2009; Ball et al. 2009).

Three studies have shown that DNA methylation level at specific CpG sites are quantitative traits that can be located by QTL mapping too, as summarized in Table 5.3. The methylation level is quantified as percentage of methylation at a specific CpG site, with values ranging from 0 to 1.

Figure 5.3 shows an example of mQTL converging with eQTL for *IRF6*. This is one of the very few examples that genotype-expression-methylation has a three-way correlation fitting the classical model of gene expression regulation.

Only the Illumina Infinium Human Methylation27 and Methylation450 arrays are available for accurate measure of DNA methylation at many CpG sites across genome. They assay 27 K and 480 K CpG sites in the genome, respectively. A study by Chen et al. discovered that about 3,000 probes in the Meth27 array may cross-hybridize to more than one genomic region, and several hundreds of probes carry SNPs (Chen et al. 2011b). We analyzed their data and identified 58 probes carrying common SNPs (MAF$\geq$0.05). A list of these "affected" probes is also provided through our website (http://bioinfo.psych.uic.edu/ArrayGenes/SNPsInProbes.jsp).

Table 5.3 Methylation QTL mapping studies

Authors	Samples	Platforms	Findings
Zhang et al. (2010)	153 cerebellum cortex, Caucasian,	Affymetrix 5.0 array for genotyping; Infinium HumanMethylation27 BeadChips	736 CpG sites showed phenotype-wide significant cis-association with 2,878 SNPs (after permutation correction for all tested markers and methylation phenotypes) Trans-associations of 12 CpG sites and 38 SNPs remained significant after phenotype-wide correction
Gibbs et al. (2010)	four human brain regions each: cerebellum, frontal cortex, temporal cortex, and pons regions from 150 individuals (600 samples total) (European descent)	Infinium HumanHap550 Beadchips; Infinium HumanMethylation27 BeadChip	7,966–12,081 cis-mQTLs, 2,893–4,653 trans-mQTLs (permutation for SNPs tested, FDR for traits tested)
Bell et al. (2011)	77 HapMap YRI cell lines.	HapMap release 27 genotype data were obtained for 3.8 M autosomal SNPs; Il lumina HumanMethylation27 DNA Analysis BeadChip	180 CpG sites in 173 genes that were associated with nearby SNPs (putatively in cis, usually within 5 kb) at a false discovery rate of 10%

Note: All these studies used methylation27 chip, which targets 27,000 CpG sites. The numbers of SNPs tested varied

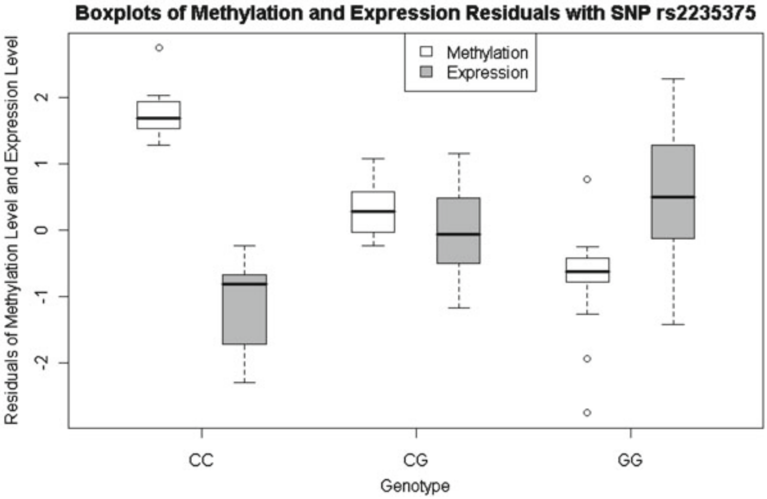

Fig. 5.3 DNA methylation and gene expression of IRF6 is correlated with genotypes of rs2235375. DNA methylation and gene expression are negatively correlated (From Zhang et al. 2010)

Besides these two Infinium BeadChips, DNA methylation level can also be accurately measured by pyrosequencing but with a much smaller throughput. Methylome sequencing is expected to provide better coverage through the genome. But the cost is still prohibitively high today for a population-based study.

It should be noted that DNA methylation may be a more complicate biological process than we expected. 5-hydroxylmethylcytosine (5hmC) was discovered to be abundant in brains (Kriaucionis and Heintz 2009) and embryonic stem cells (Tahiliani et al. 2009). The function of 5hmC remains largely unknown. It may be an intermediate step of DNA demethylation. It may have its own specific binding proteins. MeCP2 and other major methyl-CpG-binding proteins will not bind 5hmC (Valinluck et al. 2004; Jin et al. 2010).

The presence of 5hmC may interfere the measure of 5mC. Some enzymatic digestion methods and bisulfite-based methods including Infinium or pyrosequencing method cannot differentiate 5hmC from 5mC (Huang et al. 2010b). So the BeadChip results should be a measure of combined 5hmC and 5mC.

In genome-wide assessment of gene expression-DNA methylation correlation, we see many incidence of poor correlations between methylation and expression, or positive correlation. 5hmC may partially play a role in that discrepancy. Jin et al. reported that in human brain, 5hmC in gene bodies were more positively correlated with gene expression than 5mC (Jin et al. 2011). Eventually, mQTL mapping will need to be differentiated into mQTL for 5mC and hmQTL for 5hmC. But the technology is not there yet.

Another interesting observation is that mQTL and eQTL seem to be largely independent. SNPs associated with DNA methylation are not the one showing association with expression level. Very few SNPs affect both expression and CpG methylation at the same time(Gibbs et al. 2010). Possible explanations include the following: some of the genetically regulated methylations, regardless of 5mC, 5hmC difference, do not affect gene expression significantly or the correlations were not detected due to limited statistical power or those regulations were not captured by the current technology. If the methylation does not affect expression, would it likely to be functional? The answer is "yes" as it will be discussed below, which showed the mQTL SNPs were enriched in disease GWAS signals. Our hypothesis-to-be-tested today is that DNA methylation has function beyond regulating gene expression. It is known that DNA methylation is also regulating DNA stability (Lorincz et al. 2002), repressing retrotransposons (Kuhlmann et al. 2005), imprinting (Li et al. 1993). Anything else? ought to be discovered in the future. Better technology and larger sample size study will improve our understanding of regulation of both gene expression and DNA methylation.

5.3 Other Types of Quantitative Traits

Many other molecular measures, such as protein and lipid level, enzyme activity, and metabolites, can be used for QTL mapping. A few examples are summarized below.

Melzer et al. studied levels of 42 proteins in 1,200 fasting individuals for their associations with about half a million SNPs, to map protein quantitative trait loci (pQTLs). Eight cis-associations were detected with effect sizes ranging from 0.19 to 0.69 standard deviations per allele. A trans-association was observed but failed to be replicated (Melzer et al. 2008).

GWAS of plasma liver-enzyme in 12,419 individuals revealed six regulatory loci reaching genome-wide significance (Yuan et al. 2008).

Study of 363 metabolites in serum of 284 male participants did not detect association that can survive the most conservative multiple testing correction, but two loci reach genome-wide significance with $p < 4\text{e}{-8}$ (Gieger et al. 2008).

Metabolic/metabolite quantitative trait locus was also called mQTL. In a study of approximately 200 individuals for 526 metabolite traits, concentrations of four metabolites, including trimethylamine, 3-amino-isobutyrate, an N-acetylated compound, and dimethylamine, measured in urine or plasma exhibited significant and replicable QTLs (Nicholson et al. 2011). The mapped QTLs can explain 40–64% of variations**.**

GWAS mapping of lipid phenotypes in 1,087 individuals using a 100 K genotyping array failed to produce convincing result (Kathiresan et al. 2007).

Thirty-three traits and forty three matched ratios of circulating sphingolipid, including sphingomyelin (SM), dihydrosphingomyelin (Dih-SM), ceramide (Cer), and glucosylceramide (GluCer) single lipid species, were studied in European populations for 4,400 subjects. Thirty two SNPs in five distinct loci reach genome-wide significance ($p < 1\text{e}{-10}$) (Hicks et al. 2009).

5.4 Software and Algorithm for QTL Mapping

Linear regression is the most frequently used method in QTL mapping. Plink (Purcell et al. 2007) (http://pngu.mgh.harvard.edu/~purcell/plink/) is widely used for that. Other software like R/eMap (http://www.bios.unc.edu/~wsun/software/eMap.pdf) and Matrix eQTL (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/) also can do the job. Matrix eQTL claimed to have the most efficient algorithm. Most of the software provides methods for multiple testing correction.

In concern of the non-normal distribution of the data, nonparametric methods such as Spearman Rank correlation test (Montgomery et al. 2010) and Kruskal-Wallis test (Schadt et al. 2008) were also used. But the covariate and power issue may limit the use of those nonparametric methods.

A different method, VBQTL, uses a probabilistic approach for eQTLs mapping. It jointly models contributions from genotype as well as known and hidden confounding factors to achieve better power. (Stegle et al. 2010)

Microsoft Linear Mixed Models (LMM-EH-PS) (Listgarten et al. 2010), (http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/MSLMM/) uses linear mixed-effects models to model hidden confounders in association studies. It aims to control experimental batch effects and population structure and other possible confounding factors altogether. It was shown to outperform other methods including Inter-sample Correlation Emended (ICE) (Kang et al. 2008) and Surrogate Variable Analysis (SVA)(Leek and Storey 2007) for better calibrated p-values and maximum power. All these new methods need more careful comparative evaluations to find their best-fits in actual studies.

5.5 Applications of QTL Mapping in Genetic Studies of Complex Diseases

Although statistical associations between SNPs and those molecular traits reached significance level, question could still be raised: Are those QTL SNPs truly informative or directly involved in complex diseases at all? At least three studies showed that the disease-associated SNPs from GWASs are significantly more likely to be eQTL SNPs (eSNPs) than to be other random minor allele frequency (MAF)-matched SNPs from high-throughput GWAS platforms or from the HapMap(Nicolae et al. 2010; Richards et al. 2012; Gamazon et al. 2012). Signals from the NHGRI GWAS catalog were shown to be enriched for eQTLs detected in HapMap LCLs (Nicolae et al. 2010). Schizophrenia GWAS SNPs with $p<0.5$ were enriched with eSNPs detected in brain originally reported by Myers et al. (2007) and Webster et al. (Richards et al 2012) Bipolar disorder GWAS signals with $p<0.001$ or <0.0001 were all enriched with eQTL and mQTL SNPs detected in cerebellum(Gamazon et al. 2012).

GWAS of complex diseases have been restrained by the multiple testing problem when millions of SNPs are tested. If we can limit the tests to functional SNPs, number of tests may be greatly reduced. Our study using only mQTL SNPs detected in cerebellum has proved that it is a fruitful practice. A novel bipolar disorder association was discovered for SNP rs12618769, which can survive the lowered genome-wide significance threshold coming with the reduced number of tests (Gamazon et al. 2012). This association is replicated in three datasets, including the largest bipolar collection from Psychiatric Genomics Consortium (PGC, 11,974 cases and 51,792 controls) with $p=0.0031$. SNP rs12618769 is a cis-mQTL of *INPP4A*.

In a Crohn's disease (CD) study, after confirming overrepresentation of cis-eQTLs in the known CD-associated loci, association studies of eSNPs identified two likely novel risk genes: *UBE2L3* and *BCL3* for CD (Fransen et al. 2010).

Several other GWASs of psychiatric diseases have also incorporated brain eQTL data to enhance the statistical powers, leading to identification of novel risk genes. The papers are expected in 2012.

We are moving into the era of next-generation sequencing (NGS). NGS is expected to be fruitful for the purpose of complex disease association mapping. Individuals are likely to carry tens of millions of DNA variants, and testing all the variants for disease association unselectively would be a statistical nightmare, requiring impossibly large sample sizes. Limiting the studies to functional variants or the most likely relevant genes will be the optimal and probably the only choice. QTL mapping of molecular traits will be one efficient approach discovering those functional variants. Meanwhile, this need will push the QTL mapping to the use of NGS to replace SNP array as many of the variants detected in NGS cannot be tested in SNP array.

5.6 Database for QTL Mapping Results

While QTL mapping studies are blooming, several databases have been created for collecting and sharing those results. A number of databases have dedicated for sharing eQTL data, including Scandb (http://www.scandb.org/newinterface/about.html), Genevar (GENe Expression VARiation, http://www.sanger.ac.uk/resources/software/genevar/), and eQTL Browser (http://eqtl.uchicago.edu/help.html).

Scandb provides rich annotation for both SNP and gene (Gamazon et al. 2010). eQTL data used those from the HapMap data. A unique feature of this database is that it incorporates LD information among SNPs. CNV is also included.

Genevar allows researchers to investigate eQTL associations within a gene locus of interest in real time. It currently contains gene expression and genotype data from three cell types (fibroblast, LCL, and T-cell) of 75 Geneva GenCord individuals(Dimas et al. 2009) and three tissue types (166 adipose, 156 LCL and 160 skin samples) from healthy female twins of the MuTHER resource(Nica et al. 2011).

eQTL Browser collected seven eQTL datasets and provided interface similar to HapMap browser: Liver eQTL by Schadt et al. (2008); brain eQTL by Myers et al. (2007); HapMap LCL by Stranger et al. (2007), Veyrieras et al. (2008), Pickrell et al. (2010), and Montgomery et al. (2010); and monocyte eQTL by Zeller et al. (2010).

NCBI GTEx (Genotype-Tissue Expression, http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi) eQTL Browse now carries seven datasets of LCL, brain, and liver from four studies (Stranger et al. 2007; Schadt et al. 2008; Montgomery et al. 2010; Gibbs et al. 2010).

Phenotype-Genotype Integrator (PheGenI, http://www.ncbi.nlm.nih.gov/gap/PheGenI) merges NHGRI genome-wide association study (GWAS) catalog data with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, GTEx, and dbSNP.

PharmGKB (http://www.pharmgkb.org/) provide SNPs associated with drug response along with curated data of pharmacogenomics literature. Most of the data were not reviewed in this chapter.

So far, no single database integrated all the QTL mapping studies that have been published. Existing databases could be considered as good prototypes of an ideal database that can facilitate the studies of complex diseases. We hope that, with better comprehensive data integration, more risk genes of complex diseases will be discovered.

Summary, new experimental platform will ensure better coverage and more accurate measure of all the molecular traits. Larger sample size study of all the disease-relevant tissues or their proxies will be investigated for QTL mapping. These studies will provide rich functional annotation of human genetic variants. They will serve as important disease intermediate phenotypes and a venue approaching understanding of complex disease.

Acknowledgement I would like to thank Drs. Judith Badner, Yin Yao Shugart, and Chao Chen for critical readings and comments.

References

Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC. Sequence polymorphisms cause many false cis eQTLs. PLoS One. 2007;2(7):e622.

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27(4):361–8.

Baum AE, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry. 2008;13(2):197–207.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12(1):R10.

Breen G, et al. Replication of association of 3p21.1 with susceptibility to bipolar disorder but not major depression. Nat Genet. 2011;43(1):3–5.

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011a;6(2):e17238.

Chen YA, Choufani S, Ferreira JC, Grafodatskaya D, Butcher DT, Weksberg R. Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray CHEN2011. Genomics. 2011b;97(4):214–22.

Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009;10(9):595–604.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365–9.

Cichon S, et al. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. Am J Hum Genet. 2011;88(3):372–81.

Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, Wang H, Williams A, Blume JE. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol. 2007;8(4):R64.

Colantuoni C, et al. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature. 2011;478(7370):519–23.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10(3):184–94.

Damerval C, Maurice A, Josse JM, de Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics. 1994;137(1):289–301.

Deng J, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol. 2009;27(4):353–60.

Dimas AS, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325(5945):1246–50.

Dixon AL, et al. A genome-wide association study of global gene expression. Nat Genet. 2007;39(10):1202–7.

Dobrin R, Greenawalt DM, Hu G, Kemp DM, Kaplan LM, Schadt EE, Emilsson V. Dissecting cis regulation of gene expression in human metabolic tissues. PLoS One. 2011;6(8):e23480.

Duan S, Zhang W, Bleibel WK, Cox NJ, Dolan ME. SNPinProbe_1.0: a database for filtering out probes in the Affymetrix GeneChip human exon 1.0 ST array potentially affected by SNPs. Bioinformation. 2008;2(10):469–70.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.

Emilsson V, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452(7186):423–8.

Fehrmann RS, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011;7(8):e1002197.

Ferreira MA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 2008;40(9):1056–8.

Fransen K, et al. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum Mol Genet. 2010;19(17):3482–8.

Gamazon ER, Zhang W, Dolan ME, Cox NJ. Comprehensive survey of SNPs in the Affymetrix exon array using the 1000 Genomes dataset. PLoS One. 2010;5(2):e9366.

Gamazon ER, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants GAMAZON2012. Mol Psychiatry. 2012.

Gibbs JR, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6(5):e1000952.

Gieger C, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008;4(11):e1000282.

Heap GA, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet. 2010;19(1):122–34.

Heinzen EL, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. 2008;6(12):e1.

Hicks AA, et al. Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet. 2009;5(10):e1000672.

Hsiao CL, Lian I, Hsieh AR, Fann CS. Modeling expression quantitative trait loci in data combining ethnic populations. BMC Bioinformatics. 2010;11:111.

Huang J, et al. Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression HUANG2010. Am J Psychiatry. 2010a;167(10):1254–63.

Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One. 2010b;5(1):e8888.

Jin SG, Kadam S, Pfeifer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res. 2010;38(11):e125.

Jin SG, Wu X, Li AX, Pfeifer GP. Genomic mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Res. 2011;39(12):5015.

Johnson JM, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 2003;302(5653):2141–4.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.

Johnson MB, et al. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron. 2009;62(4):494–509.

Jones PA. The DNA methylation paradox. Trends Genet. 1999;15(1):34–7.

Kang HM, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 2008;180(4):1909–25.

Kang HJ, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011;478(7370):483–9.

Kathiresan S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Med Genet. 2007;8 Suppl 1:S17.

Klein RJ, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385–9.

Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science. 2009;324(5929):929–30.

Kuhlmann M, et al. Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. Nucleic Acids Res. 2005;33(19):6405–17.

Kwan T, et al. Heritability of alternative splicing in the human genome. Genome Res. 2007;17 (8):1210–18.

Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.

Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature. 1993;366 (6453):362–5.

Listgarten J, Kadie C, Schadt EE, Heckerman D. Correction for hidden confounders in the genetic analysis of gene expression. Proc Natl Acad Sci U S A. 2010;107(38):16465–70.

Liu C. Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases LIU2011. Neurosci Bull. 2011;27(2):123–33.

Liu C, Cheng L, Badner JA, Zhang D, Craig DW, Redman M, Gershon ES. Whole-genome association mapping of gene expression in the human prefrontal cortex. Mol Psychiatry. 2010;15(8):779–84.

Liu Y, et al. Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder LIU2011. Mol Psychiatry. 2011;16(1):2–4.

Lorincz MC, Schubeler D, Hutchinson SR, Dickerson DR, Groudine M. DNA methylation density influences the stability of an epigenetic imprint and Dnmt3a/b-independent de novo methylation. Mol Cell Biol. 2002;22(21):7572–80.

Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011;27(2):72–9.

Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009;461 (7265):747–53.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.

McMahon FJ, et al. Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. Nat Genet. 2010;42(2):128–31.

Melzer D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008;4(5):e1000072.

Min JL, et al. The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. PLoS One. 2011;6(7):e22070.

Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. 2001;29(13):2850–9.

Moffatt MF, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007;448(7152):470–3.

Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet. 2004;75(6):1094–105.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464(7289):773–7.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430(7001):743–7.

Myers AJ, et al. A survey of genetic human cortical gene expression. Nat Genet. 2007;39(12):1494–9.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9.

Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C. Genome-wide survey of allele-specific splicing in humans. BMC Genomics. 2008;9:265.

Nica AC, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011;7(2):e1002003.

Nicholson G, et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. PLoS Genet. 2011;7(9):e1002270.

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4):e1000888.

Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.

Powell JE, Henders AK, McRae AF, Wright MJ, Martin NG, Dermitzakis ET, Montgomery GW, Visscher PM. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res. 2011;22(3):456–66.

Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

Rauch TA, Wu X, Zhong X, Riggs AD, Pfeifer GP. A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A. 2009;106(3):671–8.

Richards AL, et al. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Mol Psychiatry. 2012;17(2):193–201.

Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6(5):e107.

Schroeder A, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3.

Schwartz D. Genetic studies on mutant enzymes in maize. III. Control of gene action in the synthesis of Ph 7.5 esterase. Genetics. 1962;47(11):1609–15.

Sklar P, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011;43(10):977–83.

Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770.

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. Am J Hum Genet. 2007;80(3):502–9.

Stranger BE, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005;1(6):e78.

Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848–53.

Tahiliani M, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009;324(5929):930–5.

Valinluck V, Tsai HH, Rogstad DK, Burdzy A, Bird A, Sowers LC. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). Nucleic Acids Res. 2004;32(14):4100–8.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008;4(10):e1000214.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.

Wang KS, Liu XF, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophr Res. 2010;124(1–3):192–9.

Webster JA, et al. Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet. 2009;84(4):445–58.

Wheeler HE, et al. Sequential use of transcriptional profiling, expression quantitative trait mapping, and gene association implicates MMP20 in human kidney aging. PLoS Genet. 2009;5(10):e1000685.

Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. 2002;30(17):3754–66.

Yuan X, et al. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. Am J Hum Genet. 2008;83(4):520–8.

Zeller T, et al. Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010;5(5):e10693.

Zhang W, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet. 2008;82(3):631–40.

Zhang D, et al. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet. 2010;86(3):411–19.

Chapter 6
Renewed Interest in Haplotype: From Genetic Marker to Gene Prediction

Shuying Sue Li, Xinyi Cindy Zhang, and Lue Ping Zhao

Abstract A haplotype refers to a set of multiple SNP alleles from one parental genome and has a clear genetic interpretation, often as a biologically meaningful quantity. With the current biotechnologies for genotyping diploidic human genomes (a pair of haplotypes), however, genotype data include only partial haplotype information and, in general, are insufficient to directly infer a pair of haplotypes (with more than two SNP loci) since SNP genotypes are typed locus-by-locus. Using genotype data, haplotype analysis methods refer to a class of statistical genetic analysis methods, for inferring haplotypes, or estimating haplotype frequencies (and related statistics), or assessing haplotypic associations with a phenotype. These methods are an important set of statistical tools for genetic analyses. With advent of both genotyping/sequencing technologies, we anticipate an increasing interest in haplotype-based association analysis. In this chapter, we have introduced the concept of haplotype and its roles in genetic studies, have also documented the aspects of earlier method developments, have described some key methods and related software,

S.S. Li
Statistical Center for HIV/AIDS Research & Prevention (SCHARP),
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer
Research Center, Seattle, WA 98109, USA
e-mail: sli@fhcrc.org

X.C. Zhang
Biostatistics Program, Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
e-mail: xzhan2@fhcrc.org

L.P. Zhao (✉)
Biostatistics Program, Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Department of Biostatistics and Department of Epidemiology,
School of Public Health and Community Medicine,
University of Washington, Seattle, WA 98109, USA
e-mail: lzhao@fhcrc.org

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_6, © Springer Science+Business Media Dordrecht 2012

and have discussed the renewed interest in recent years. While exhaustively reviewing literature on haplotype analysis is not of primary interest, this chapter serves an introduction to the haplotype analysis, opening a door to a rich and dynamic set of literature accumulated in the past decades.

Keywords Haplotype • Haplotype inference • Haplotype-based association • Gene prediction

6.1 Introduction

Human genome consists of 22 autosomal chromosome pairs and one sex chromosome pair. One member of each chromosome pair is inherited from a person's father; the other member of the pair is inherited from that person's mother. The whole genome includes approximately ~20–25 thousand genes and 3.2 billion nucleotide base pairs (Bentley et al. 2001; Venter et al. 2001). The sequences of nucleotides from the same chromosome, either paternal or maternal chromosome, are called haplotypes. Haplotypes dictate the RNA transcriptions and hence proteins and therefore are considered as functional units. Many genes have been identified to be associated with diseases (King et al. 2002). In the previous association studies, different types of genetic markers were used to map disease-associated gene, such as the restricted length polymorphisms (RFLP) (Lander and Botstein 1986), short tandem repeats (STR) (Gyapay et al. 1994), microsatellite (NIH/CEPH Collaborative Mapping Group 1992; Murray et al. 1994), human leukocyte antigen (HLA) alleles, and single nucleotide polymorphisms (SNPs) (Wang et al. 1998; Zhao et al. 1998). The statistical methods developed for analyses include the linkage analyses for mapping disease traits (Zhao et al. 1998) and association analyses for fine-mapping (Xiong and Guo 1997) and characterizing genetic functions (Collins and Morton 1998). Due to the breathtaking development of SNP array technologies in recent years, SNPs are now commonly used as genetic markers in the recent either candidate gene association studies or genome-wide association studies (GWAS).

For candidate gene association studies, multiple SNPs within a candidate gene are typically genotyped. The most commonly used technologies generate unphased SNP genotypes (without knowing the parental origin of the alleles). Therefore, haplotypes would not be directly obtained from the genotyping. Although some genotyping technologies can generate haplotypes directly by dissecting out single chromosomes (Green et al. 1998; Patil et al. 2001), such technologies are experimentally challenging and cost prohibitive for the use in human research. One way to construct individual haplotypes from unphased genotypes is to genotype for other family members of that person (Wijsman 1987), which is considered as expensive and unpractical. A practical option is to statistically infer individual haplotypes from unphased genotypes. A latter option has been used widely in the genetic research. Clark's heuristic algorithm was probably among the first statistical methods for inferring haplotypes from genotypes of unrelated individuals (Clark 1990).

Several maximum likelihood methods were developed thereafter (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995; Stephens et al. 2001; Lin et al. 2002; Niu et al. 2002; Qin et al. 2002; Li et al. 2003; Stephens and Donnelly 2003). Many of those methods successfully and efficiently inferred haplotypes and were implemented in various different software packages. Their inference for haplotypes was evaluated (Li et al. 2007).

There are also haplotype-based association methods developed to correlate the candidate genes with the disease phenotype (Epstein and Satten 2003; Stram et al. 2003; Zhao et al. 2003; Chen et al. 2004; Lin 2004; Schaid 2004; Spinka et al. 2005; Lin and Zeng 2006; Tzeng et al. 2006; Zeng et al. 2006).

In anticipation of GWAS, major research efforts were placed in identifying haplotype blocks genome-wide (Daly et al. 2001; Goldstein 2001; Patil et al. 2001; Gabriel et al. 2002; Cardon and Abecasis 2003; Altshuler et al. 2005). However, while GWAS have been routinely conducted, few have published results based on haplotype associations and most studies have focused on SNP associations instead. Two reasons might be for the use of SNP-based instead of haplotype-based analyses in conducting GWAS analysis. First, SNP-based association analysis is easier to analyze and the results tend to be more stable than those from the haplotype analysis (due to reconstructing haplotype structures from multiple unphased SNPs). Secondly and more importantly, there is not universally agreed haplotype block definition and the blocks are likely different between the study populations.

So far, GWAS have identified at least 2,000 SNPs that are associated with common diseases and related traits (http://www.genome.gov/gwastudies). Based on these SNPs hits, researchers desire to identify the functional units (haplotypes) that contribute to the genetic associations. It is notable from those findings that several SNP hits were in the human major histocompatibility complex (MHC) region (Larsen and Alper 2004; Sabeti et al. 2007; The Wellcome Trust Case Control Consortium 2007; Asano et al. 2009; Hirschfield et al. 2009; Stefansson et al. 2009; Tse et al. 2009; Reveille et al. 2010; Fellay et al. 2011). Many HLA alleles within MHC had long been known as associations with many infectious, autoimmune diseases, and other immune-related diseases. To bridge the new and old findings, there is a need to link the SNPs with HLA alleles. HLA alleles can be considered as haplotypes of variants within 1 or 2 hypervariable regions in the HLA gene region. Based on their linkage equilibrium (LD) with the SNPs within the flanking region of the HLA genes, several methodologies were developed to predict HLA alleles using SNP genotypes (Leslie et al. 2008; Li et al. 2011; Zhang et al. 2011).

With the current advent of high throughput sequencing technologies, the 1000 Genomes Project Pilot 3 had completed sequencing of targeted regions, including 906 genes and 8,140 exons, of 697 healthy subjects and saved in a database at a public domain. The full sequences of those genes or exons will provide the research community a great opportunity to reanalyze the GWAS data to study the gene associations with the diseases. Anticipating for that effort, we constructed the models that predict the gene alleles (like HLA alleles) from the SNPs within and flank the genes using the methods described in (Li et al. 2011; Zhang et al. 2011). The details are presented elsewhere. The research community can use these models to predict the gene alleles from their GWAS data and correlate the gene alleles with the disease phenotypes for associations.

The chapter is organized as the following: the haplotype inference methods and their comparisons are presented in Sect. 6.2, the haplotype-based association methods and their comparisons are in Sect. 6.3, and the methods of predicting gene alleles using SNP genotypes are in Sect. 6.4. This chapter intends to help readers become familiar with some haplotype analysis history and does not strive to be an exhaustive review of all methods. We hope this chapter will help the readers in selecting the appropriate methods for their analyses. We have developed the software to perform the analyses discussed in this chapter. The website for the software download is provided at the end of this chapter.

6.2 Haplotype Inference

A key component in all haplotype-based analyses is to infer haplotypes from unphased genotypes. In this section, we start with describing the statistical framework and the methods for haplotype inference.

Consider a sample of n unrelated individuals from a study population. From each individual sample, we select q SNP loci in a specific region, e.g., a candidate gene, for genotypes. The genotype at each locus consists of two alleles, one from the maternal chromosome and another from the paternal chromosome. The parental origin of alleles is called phase. Since the phase of genotypes is generally not available, the haplotypes have to be inferred from the unphased genotypes. Let $G_i = \left(g_{i1},\ldots,g_{iq}\right)$ denote the genotypes of the q SNPs for the i-th individual. When a genotype is heterozygous (consists of two different alleles), there are two resolutions for phase. If G_i is heterozygous at m loci, there are 2^m possible resolutions for phase which results in 2^{m-1} distinct pairs of haplotypes. Each possible pair of haplotypes is associated with a probability that should be a function of haplotype frequencies in the study population. Suppose there are T distinct haplotypes in the population. Let $\pi = \left(\pi_1,\pi_2,\ldots,\pi_T\right)$ denote the frequencies of the T haplotypes. The distribution of haplotypes is assumed to follow a multinomial distribution with the parameter π. Let $P_i = \left(p_{i1},p_{i2},\ldots,p_{iq}\right)$ denote the phase of G_i with the phase of the first heterozygous fixed. Given phases P_i, genotypes G_i uniquely determine a pair of haplotypes (diplotype) $\left(\dot{h}_i,\ddot{h}_i\right)$, i.e., $G_i \mid P_i = \left(\dot{h}_i,\ddot{h}_i\right)$. The likelihood function of genotypes given the haplotype frequencies may be written as

$$L(\pi) = \prod_{i=1}^{n} f(G_i \mid \pi) = \prod_{i=1}^{n} \sum_{P_i} f(G_i \mid P_i,\pi) f(P_i) = \prod_{i=1}^{n} \sum_{P_i} f(\dot{h}_i,\ddot{h}_i \mid \pi) f(P_i), \quad (6.1)$$

where $f(\dot{h}_i,\ddot{h}_i \mid \pi) = f(\dot{h}_i \mid \pi) f(\ddot{h}_i \mid \pi)$ under Hardy-Weinberg Equilibrium (HWE), $f(\dot{h}_i \mid \pi)$ and $f(\ddot{h}_i \mid \pi)$ are the probabilities of haplotypes $\dot{h}_i$ and $\ddot{h}_i$ given the population's haplotype frequencies π, and $f(P_i)$ is the prior probability of phase. The estimate of π is obtained from maximizing the likelihood (6.1).

The posterior distribution of phase, which determines the probability of diplotype, is then estimated from the estimate of haplotype frequency π by

$$f(P_i \mid G_i) = \frac{f(G_i \mid P_i;\pi) f(P_i)}{\sum_{P_i} f(G_i \mid P_i;\pi) f(P_i)} = \frac{\pi_{\dot{h}_i}\pi_{\ddot{h}_i}}{\sum_{P_i:G_i \mid P_i=(\dot{h}_i,\ddot{h}_i)} \pi_{\dot{h}_i}\pi_{\ddot{h}_i}}.$$

Different methods make different assumptions for the prior distribution $f(P_i)$ and employed different algorithms to maximize the likelihood function in (6.1). The probability of phase $f(P_i)$ is assumed to be constant in all the empirical methods including expectation-maximization (EM) and equating equation (EE) methods. Excoffier and Slatkin used the expectation-maximization (EM) algorithm to obtain the maximum likelihood (ML) estimators of π and the bootstrap method to estimate the standard errors of estimators of π (Excoffier and Slatkin 1995). When the number of SNP genotypes is large, the computation for ML estimators becomes a great burden. Qin et al. utilized the partition ligation computational strategy along with EM algorithm to infer haplotypes with large numbers of SNP loci (implemented in PL-EM) (Qin et al. 2002). We applied the estimation equation technique and the partition strategy to effectively infer haplotypes with large data sets and to effectively estimate the frequencies of haplotypes and standard errors of the estimated haplotype frequencies (Li et al. 2003). Our method was implemented in HPlus.

Unlike the empirical approaches described above, the Bayesian methods assume a prior distribution of phase which is not constant. For example, Stephens and Donnelly assumed that the haplotypes were coalescent during the evolution process and modeled for the mutations and the recombination rates (Stephens and Donnelly 2003). The model, called a coalescent-based model, has been the basis of many population-based statistical phasing methods, including PHASE (Li and Stephens 2003), fastPHASE (Scheet and Stephens 2006), MACH (Li et al. 2010), and IMPUTE2 (Howie et al. 2009). PHASE was considered as the most accurate method but less computationally efficient among the four programs (Browning and Browning 2011). MACH and IMPUTE2 were mostly used for imputing the genotypes at the un-genotyped loci.

Another Bayesian approach assumed a prior distribution of haplotype frequency π that follows a Dirichlet distribution with hyper-parameter $\beta = \left(\beta_1,\dots,\beta_T\right)$ (Niu et al. 2002). Using Gibbs sampling algorithm, the authors sampled a pair of compatible haplotypes for each individual and estimated the haplotype frequencies. This method was implemented in HAPLOTYPER.

Li et al. carried out the comparisons of the performances between two empirical methods (PL-EM as an EM method and HPlus as an EE method) and two Bayesian methods (PHASE as a coalescent-based method and HAPLOTYPER as a GIBB sampling method) based on X-chromosome data from HapMap project and simulations (Li et al. 2007). Each method yields the estimation of haplotype frequencies and the prediction for the individuals' diplotypes. The accuracy of the estimation of haplotype frequencies from each method was evaluated by the similarity index defined as $I(\hat{\pi};\pi) = 1 - 0.5\sum_{j=1}^{T} \mid \hat{\pi}_j - \pi_j \mid$, where π_j and $\hat{\pi}_j$ are the true and the estimated

Table 6.1 Performances of haplotype inference methods

	Empirical method		Bayesian method	
Performances	PL-EM	HPlus	PHASE	HAPLOTYPER
Similarity index				
Mean	0.989	0.990	0.986	0.991
Median	1.0	1.0	1.0	1.0
Standard deviation	0.029	0.024	0.040	0.024
Range	(0.733, 1.0)	(0.833, 1.0)	(0.292,1.0)	(0.733,1.0)
Prediction rate				
Mean	0.989	0.990	0.990	0.991
Median	1.0	1.0	1.0	1.0
Standard deviation	0.029	0.025	0.040	0.025
Range	(0.733, 1.0)	(0.833, 1.0)	(0.283, 1.0)	(0.733, 1.0)

frequency of the j-th haplotype. The value of the similarity index ranges from zero to one corresponding to the accuracy that ranged from 0 to 100%. The accuracy of the prediction for the individuals' diplotypes was measured by the percent of correct prediction. The simulations showed that all four methods yielded very similar and accurate results using both measures. Table 6.1 shows the comparison results based on the X-chromosome data of the mothers in HapMap trios. Given the parent-child trio X-chromosome data, the phases of the mothers' genotypes on the X-chromosome could be resolved completely; therefore, the mothers' true haplotypes on the X-chromosome became known. The whole X-chromosome was divided into several hundreds of haplotype blocks. Among these, 500 blocks were randomly selected and inferred for their haplotype pairs using each of the four methods. The results were compared with the true haplotypes. The comparisons showed that all four programs yielded 99% or greater accuracy in both haplotype frequency estimation and diplotype prediction averagely over 500 blocks (Table 6.1). HPlus gave the narrowest range [0.83, 1.0] and PHASE gave the widest range [0.29, 1.0] in both accuracy measures. Moreover, HPlus was more than 100 times computationally efficient than PHASE. PHASE probably produced more accurate estimates for rare haplotypes than other methods, under appropriate assumption for the corresponding population genetic model.

6.3 Haplotype-Based Association

Below, we first introduce the models and estimations and then present a comparison of the estimation results of the prospective and retrospective approaches using simulations under various models of haplotype distribution.

6.3.1 Models and Estimations

Let y_i denote the phenotype and X_i denote the environmental covariates (such as subjects' characteristics and clinical variables). The phenotype can be a quantitative (continuous) measure (e.g., blood pressure) or a qualitative (binary) measure (e.g., disease status in a case-control study) or an on-set time (e.g., the time to develop a disease in a cohort study). Suppose a primary interest of the study is to correlate the gene with the phenotype and to adjust for environmental covariates. For a continuous or binary phenotype, the log-likelihood function based on the observed data then may be written as

$$l(\beta,\pi)=\sum_{i=1}^{n}\log f(y_i \mid G_i,X_i)=\sum_{i=1}^{n}\sum_{(\dot{h}_i,\ddot{h}_i)\in S(G_i)}\log f(y_i \mid (\dot{h}_i,\ddot{h}_i),X_i)f(\dot{h}_i,\ddot{h}_i \mid G_i,X_i), \quad (6.2)$$

where β is a vector of the association parameters of haplotypes and other covariates with the phenotype and π is a vector of haplotype frequencies; $S(G_i)$ represents a set of all possible pairs of haplotypes (diplotypes) that give arise to the genotypes G_i ; $f((\dot{h}_i,\ddot{h}_i) \mid G_i,X_i)$ is the distribution of a haplotype pair given the genotype and covariates; and $f(y_i \mid (\dot{h}_i,\ddot{h}_i),X_i)$ is a penetrance function of the phenotype with genetic and other covariates. The design matrix of genetic part is determined by the hypothesis test of interest and a genetic model, such as recessive, dominant, or additive model. If one wants to test the effect of a specific haplotype h in contrast to all other haplotypes, the design matrix for the i-th individual with a haplotype pair $(\dot{h}_i,\ddot{h}_i)$ is an indicator function $I(\dot{h}_i = h \,\&\, \ddot{h}_i = h)$ under a recessive model, $I(\dot{h}_i = h \mid \ddot{h}_i = h)$ under a dominant model, and $I(\dot{h}_i = h)+I(\ddot{h}_i = h)$ under an additive model. Another hypothesis test of interest is to test the joint effects of multiple haplotypes in contrast to a reference haplotype. In general, the most common haplotype is treated as the reference. To be specific, suppose there are m possible haplotypes, $h_1,h_2,\ldots,h_m$ in descending order according to their frequencies $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_m$. The design matrix of genetic part is a vector of m-1 indicator functions, $(I(\dot{h}_i = h_2)+I(\ddot{h}_i = h_2), I(\dot{h}_i = h_3)+I(\ddot{h}_i = h_3),\ldots,I(\dot{h}_i = h_m)+I(\ddot{h}_i = h_m))$ for haplotypic analysis and $(I(\dot{h}_i = h_2 \,\&\, \ddot{h}_i = h_2), I(\dot{h}_i = h_2 \,\&\, \ddot{h}_i = h_3),\ldots,I(\dot{h}_i = h_m \ddot{h}_i = h_m))$ for diplotypic analysis. For rare haplotypes/diplotypes, the associated sample sizes prohibit direct assessments. One can either combine all rare haplotypes/diplotypes into a compound haplotype/diplotype or group them to the reference haplotype/diplotype.

The parameters (β, π) can be estimated using the estimating equation:

$$\begin{pmatrix} U(\beta) \\ U(\pi) \end{pmatrix} = \sum_{i=1}^{n} \sum_{(\dot{h}_i,\ddot{h}_i)\in S(G_i)} \begin{pmatrix} Z_i(y_i-\mu_i)f((\dot{h}_i,\ddot{h}_i) \mid y_i,G_i,X_i) \\ (K(\dot{h}_i)+K(\ddot{h}_i)-2\pi)f((\dot{h}_i,\ddot{h}_i) \mid y_i,G_i,X_i) \end{pmatrix}, \quad (6.3)$$

where Z_i is the design matrix, K is a vector of indicator function of haplotypes, and

$$f((\dot{h}_i,\ddot{h}_i) \mid y_i,G_i,X_i)=\frac{f(y_i \mid (\dot{h}_i,\ddot{h}_i),X_i)f(\dot{h}_i)f(\ddot{h}_i)}{\sum_{(\dot{h}_i,\ddot{h}_i)\in S(G_i)} f(y_i \mid (\dot{h}_i,\ddot{h}_i),X_i)f(\dot{h}_i)f(\ddot{h}_i)}$$

under HWE is the posterior probability of $(\dot{h}_i, \ddot{h}_i)$ given the phenotype and covariates. The standard error of the estimate is obtained using the sandwich estimation from a product of the information matrix and the expectation of estimating equation.

For a case-control study, Prentice and Pyke (1979) showed that the prospective maximum likelihood estimation of (6.2) yields valid results even though data is retrospectively ascertained.

However, the assumption of HWE might not hold in some cases. The violation of the assumption for HWE in the estimation would lead to estimation bias. The case-control design is mostly used for studying rare diseases. With the assumption of a rare disease, the posterior probability $f\left((\dot{h}_i, \ddot{h}_i) \middle| y_i, G_i, X_i\right)$ in (6.3) approximates to $\left\{\exp(y_i \beta Z_i) f(\dot{h}_i) f(\ddot{h}_i)\right\} / \left\{\sum_{(\dot{h}_i, \ddot{h}_i) \in S(G_i)} \exp(y_i \beta Z_i) f(\dot{h}_i) f(\ddot{h}_i)\right\}$. The details of the derivation are referred to (Zhao et al. 2003). With this approximation, the estimating equation for π depends on the genotype data from controls only. This modification potentially costs estimation efficiency for estimating association parameter, β. Another approach for analyzing the case-control data is the retrospective maximum likelihood method that models the distributions of haplotypes and covariates and conditions on case and control status. Like the prospective analysis, the retrospective analysis assumes HWE for haplotypes. In general, a retrospective approach is more efficient than a prospective approach under the model assumption. However, the retrospective estimates are subject to the model assumption. The violation of model assumption causes bias in the estimation results. Instead, a prospective approach is robust to the model assumption, which in this case is the HWE assumption made for the true haplotypes. Later in this section, we present a comparison of the estimation results of the two approaches using simulations under various models of haplotype distribution.

In a cohort study, the study subjects are usually observed for the development of a particular phenotype. The phenotype on-set time is recorded for those who develop a phenotype and the observation time is recorded for those who have not developed the phenotype during the observation period, which is referred as a censor time. Let T_i and C_i denote for the phenotype on-set time and the censor time, respectively. The observed data can be denoted as $(y_i, \Delta_i, G_i, X_i), i = 1, 2, \ldots, n$, where $y_i = \min(T_i, C_i)$, and $\Delta_i = I(T_i \le C_i)$. According to Cox proportional hazard model, the hazard function given the i-th individual's pair of haplotypes $(\dot{h}_i, \ddot{h}_i)$ and the environmental covariates X_i is

$$\lambda\left(t \middle| (\dot{h}_i, \ddot{h}_i), X_i\right) = \lambda_0(t) e^{\beta Z_i}, \tag{6.4}$$

where Z_i is the design matrix of the haplotypes in the formations described above and the covariates X_i. The likelihood function given the observed data is proportional to

$$L_p(\beta, \pi) = \prod_i \sum_{(\dot{h}_i, \ddot{h}_i) \in S(G_i)} \left(\Lambda_0 \{y_i\} e^{\beta Z_i}\right)^{\Delta_i} f\left((\dot{h}_i, \ddot{h}_i) \middle| X_i\right) \tag{6.5}$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function, it can be estimated using Breslow estimator. The association parameters β and the haplotype frequencies π were estimated jointly in (Lin 2004). However, since the haplotype frequencies do not depend on the association parameters and environmental covariates, we recommend estimating π and β separately. After replacing π by its estimate, the estimation of β is then carried using the Lin's EM algorithm. In the $(k+1)$st iteration of the maximization step (M-step), the estimator $\hat{\beta}^{(k+1)}$ is the solution of the estimating function

$$U^k(\beta)=\sum_{i=1}^{n} \Delta_i\{E(Z_i \mid G_i,X_i)-\frac{\sum_{j=1}^{n} I(y_j \ge y_i)E(e^{\beta^k Z_i}Z_i \mid G_i,X_i)}{\sum_{j=1}^{n} I(y_j \ge y_i)E(e^{\beta^k Z_i} \mid G_i,X_i)}, \tag{6.6}$$

where $E(\cdot \mid G_i,X_i)$ is the expectation with respect to the posterior distribution of diplotypes given genotypes G_i and covariates X_i. Then the corresponding estimator of $\Lambda_0(t)$ is updated by

$$\hat{\Lambda}_0^{(k+1)}(t)=\sum_{i=1}^{n}\frac{I(y_i \le t)\Delta_i}{\sum_{j=1}^{n} I(y_j \ge y_i)E(e^{\hat{\beta}^k Z_i} \mid G_i,X_i)}. \tag{6.7}$$

In E-step, we update the posterior distribution of diplotypes given genotypes and covariates using the following formula:

$$f(\dot{h}_i,\ddot{h}_i \mid y_i,G_i,X_i)=\frac{I((\dot{h}_i,\ddot{h}_i)\in S(G_i))\exp\{\Delta_i\beta Z_i - e^{\beta Z_i}\Lambda_0(y_i)\}\hat{\pi}_{\dot{h}_i}\hat{\pi}_{\ddot{h}_i}}{\sum_{(\dot{h}_i,\ddot{h}_i)\in S(G_i)} \exp\{\Delta_i\beta Z_i - e^{\beta Z_i}\Lambda_0(y_i)\}\hat{\pi}_{\dot{h}_i}\hat{\pi}_{\ddot{h}_i}}. \tag{6.8}$$

Our simulations showed that this modified estimation procedure not only estimated β as efficiently as Lin's but also improved the convergence of estimations of both parameters (not shown).

6.3.2 *Comparison of Prospective and Retrospective Analysis for Haplotype Association in Case-Control Studies*

Satten and Epstein (2004) conducted extensive simulations for comparison of prospective and retrospective analysis for haplotype association using simulations. The genotype data was simulated based on five SNPs on chromosome 22 from the Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus (FUSION) Genetics Study (Valle et al. 1998). The haplotype frequency estimated in the FUSION data is given in Table 6.2. The genotypes were simulated by randomly drawing a pair of haplotype according to the following distribution:

$$f(h_i,h_j)=\begin{cases} F\pi_i+(1-F)\pi_i^2 & \text{if } i=j, i,j=1,\ldots,m \\ 2(1-F)\pi_i\pi_j & \text{if } i\ne j \end{cases}, \tag{6.9}$$

Table 6.2 Haplotypes of the five SNPs and their frequencies used in the simulation study (obtained from the FUSION study)

Haplotype	Frequency
10010	0.3327
01100	**0.2489**
11011	0.1416
01011	0.1409
10100	0.0611
10110	0.0336
01111	0.0129
11100	0.0101
00010	0.0063
00100	0.0037
01101	0.0035
00110	0.001
10000	0.0009
11110	0.0009
01110	0.0007
11111	0.0007
01101	0.0005

where $\pi_1, \pi_2, \ldots, \pi_m$ are the haplotype frequencies given in Table 6.2 and F is a fixation index. HWE assumption holds only when F equals to zero. HWE assumption does not hold for any other value of F. The prospective analysis was done using Schaid method (Schaid 2004) and Zhao et al. method (Zhao et al. 2003) and the retrospective analysis was done using Epstein and Satten method (Epstein and Satten 2003). The conclusions of the simulation were that (1) the prospective and the retrospective approach were comparable when assuming the haplotype effect on disease followed an additive model; (2) for dominant and recessive models of haplotype effect, the retrospective approach was more efficient than the prospective approach; and (3) with respect to the retrospective approach, the prospective approach was more robust to the departure from the model assumption HWE. Satten and Epstein (2004) concluded that the bias of the retrospective analysis dramatically reduced if the fixation index as a model parameter was estimated in the retrospective analysis.

The distribution in (6.8) is just another assumption for the distribution of haplotype pairs. In reality, we do not know the true distribution, especial for the population mixed with cases and controls. The distribution can be in another form that is different from the one in (6.8). Prentice and Zhao (1991) gave a general model for any multivariate distribution of continuous and discrete random variables. For haplotypes, the general model is in the following quadric form:

$$f(H, \dot{H}) = \Delta^{-1} exp[\theta'(H + \dot{H}) + H' \mathit{\Phi} \dot{H}], \tag{6.10}$$

where $H = (h_1, \ldots, h_m)$ is the set of all possible haplotypes, $\mathit{\Delta}$ is the normalizing constant, the canonical parameter θ is a function of haplotype frequencies, and $\mathit{\Phi}$ quantifies the departure from the HWE.

To evaluate the bias and the power of the prospective and the retrospective analysis for the SNP data generated from the distribution that is different from (6.8) like the distribution in the model in (6.9), we conducted our simulations. The prospective analysis in our simulations was done using the Zhao et al. method (Zhao et al. 2003) and the retrospective analysis using the Lin and Zeng's method (Lin and Zeng 2006), an extension of the Epstein and Satten (2003) method by including covariates. First, we repeated Satten and Epstein's simulations (Epstein and Satten 2003) for choosing $F=0$, 0.2, 0.4, and 0.6 and the parameter of the second haplotype "01100" listed in Table 6.2, $\beta_2=0$ under the null hypothesis and $\beta_2=\log(1.5)$, log(1.2) under alternative hypothesis for a recessive model and a dominant/additive model, respectively. For each configuration of simulation parameters, we generated 500 cases and 500 controls and repeated the simulation 5000 times. The summary of simulation results is presented in Table 6.3. In general, our simulation results mostly agree with the results from Satten and Epstein (Epstein and Satten 2003). Under HWD (departure from HWE), if the parameter fixation index was estimated in the retrospective analysis, the bias reduced but the efficiency over the prospective analysis reduced as well, especially for a dominant model. We experienced some non-convergence in the retrospective analysis. The number of convergences out of 5,000 simulations is reported under the last column under each analysis method in Table 6.3. We also evaluated the joint haplotypes association estimation. The conclusion is similar to the one for the specific haplotype effect under additive model.

We then generated the genotype data from the distribution of (6.9) with $\Phi_{23}=\Phi_{32}=\varphi, \varphi=-1.0,-0.5,0.5,1.0$ and the rest of elements in $\boldsymbol{\Phi}$ to be zero, which indicates that the pair of the 2nd haplotype and the 3rd haplotype is deviated from HWE. The summary of these simulations is presented in Table 6.4. The prospective analysis continued yielding unbiased estimate for any genetic effect models, either recessive, or dominant, or additive model. The retrospective analysis yielded biased estimates under the recessive and dominant model. Estimating the fixation index in the retrospective analysis did not reduce the bias since the distribution of haplotypes was in a different model from the fixation model. Interestingly, the bias of the retrospective analysis under the additive model was comparable to the one of prospective analysis, but the retrospective analysis and the prospective analysis had similar power.

These simulations showed that the retrospective analysis yielded higher power than the prospective analysis under the recessive and dominant model if the assumption of HWE held. But it would yield false positive findings if the assumption of HWE did not hold. Instead, the prospective analysis was robust to the assumption of HWE.

6.4 Gene Prediction

The current SNP genotyping platforms have several hundred SNP probes within or in the flanking region of each HLA locus. Although it was found that some common HLA alleles could be tagged by a single or multiple SNPs, this is not the case for

Table 6.3 Comparison of the haplotype association estimation results from the prospective analysis and the retrospective analysis in case-control studies (genotype data generated under fixation model with $n = 500$ for each control and case group)

			Prospective analysis				
Model	log OR(β_G)	F	Bias	SE	95% coverage	Size/power	Of converges
Recessive	log(1.0)	0.0	0.013	0.276	0.952	0.048	5,000
		0.2	–0.013	0.219	0.951	0.049	5,000
		0.4	0.013	0.191	0.946	0.054	5,000
		0.6	–0.016	0.170	0.953	0.047	5,000
	log(1.5)	0.0	0.055	0.264	0.944	0.437	5,000
		0.2	0.006	0.212	0.949	0.509	5,000
		0.4	0.016	0.183	0.947	0.648	5,000
		0.6	0.013	0.166	0.951	0.706	5,000
Dominant	log(1.0)	0.0	0.009	0.131	0.944	0.056	5,000
		0.2	–0.011	0.130	0.952	0.048	5,000
		0.4	0.009	0.134	0.950	0.050	5,000
		0.6	–0.009	0.137	0.952	0.048	5,000
	log(1.2)	0.0	0.009	0.133	0.946	0.308	5,000
		0.2	0.000	0.130	0.955	0.276	5,000
		0.4	0.008	0.135	0.947	0.298	5,000
		0.6	–0.017	0.139	0.943	0.225	5,000
Additive	log(1.0)	0.0	0.008	0.106	0.946	0.054	5,000
		0.2	–0.009	0.096	0.954	0.046	5,000
		0.4	0.007	0.090	0.953	0.047	5,000
		0.6	–0.007	0.083	0.950	0.050	5,000
	log(1.2)	0.0	–0.005	0.106	0.947	0.382	5,000
		0.2	0.005	0.098	0.942	0.499	5,000
		0.4	0.019	0.087	0.949	0.639	5,000
		0.6	0.001	0.084	0.949	0.606	5,000
Joint haplotype	log(1.0)	0.0	0.012	0.120	0.953	0.047	5,000
		0.2	–0.009	0.109	0.946	0.054	5,000
		0.4	0.006	0.101	0.949	0.051	5,000
		0.6	–0.010	0.091	0.954	0.046	5,000
	log(1.2)	0.0	–0.005	0.120	0.954	0.308	5,000
		0.2	0.009	0.110	0.949	0.422	5,000
		0.4	0.027	0.098	0.945	0.562	5,000
		0.6	0.004	0.093	0.949	0.519	5,000

Retrospective analysis									
Under HWE					Under HWD				
Bias	SE	95% coverage	Size/ power	Of converges	Bias	SE	95% coverage	Size/ power	Of converges
–0.011	0.216	0.950	0.050	4,907					
0.663	0.185	0.067	0.933	4,931	–0.019	0.181	0.950	0.050	4,911
1.229	0.177	0.000	1.000	4,939	0.009	0.163	0.947	0.053	4,894
1.698	0.173	0.000	1.000	4,961	–0.022	0.150	0.951	0.049	4,901
0.020	0.191	0.946	0.615	4,910					
0.639	0.173	0.054	1.000	4,922	–0.037	0.169	0.949	0.589	4,905
1.197	0.159	0.000	1.000	4,589	0.001	0.151	0.953	0.768	4,922
1.598	0.140	0.000	1.000	2,778	–0.001	0.144	0.953	0.795	4,919
0.006	0.122	0.948	0.052	4,905					
–0.281	0.122	0.356	0.644	4,946	–0.012	0.128	0.952	0.048	4,912
–0.536	0.123	0.008	0.992	4,957	0.006	0.132	0.949	0.051	4,893
–0.826	0.123	0.000	1.000	4,977	–0.010	0.137	0.953	0.047	4,899
0.014	0.123	0.942	0.359	4,891					
–0.285	0.120	0.347	0.131	4,931	–0.002	0.128	0.954	0.280	4,898
–0.565	0.123	0.003	0.878	4,978	0.008	0.134	0.949	0.306	4,936
–0.884	0.123	0.000	1.000	4,986	–0.018	0.138	0.945	0.227	4,925
0.007	0.106	0.946	0.054	4,906					
–0.011	0.115	0.925	0.075	4,942	–0.009	0.095	0.952	0.048	4,911
0.008	0.125	0.899	0.101	4,954	0.007	0.088	0.949	0.051	4,894
–0.013	0.132	0.880	0.12	4,977	–0.009	0.081	0.951	0.049	4,900
–0.006	0.104	0.948	0.387	4,903					
0.042	0.117	0.897	0.575	4,935	0.001	0.095	0.943	0.507	4,899
0.102	0.121	0.796	0.754	4,960	0.014	0.084	0.947	0.643	4,919
0.113	0.133	0.739	0.755	4,991	–0.003	0.080	0.952	0.618	4,931
0.009	0.116	0.953	0.047	4,902					
–0.012	0.128	0.919	0.081	4,949	–0.01	0.106	0.947	0.053	4,917
0.008	0.139	0.902	0.098	4,952	0.007	0.098	0.949	0.051	4,899
–0.016	0.146	0.877	0.123	4,974	–0.011	0.089	0.952	0.048	4,900
–0.008	0.115	0.949	0.319	4,899					
0.047	0.128	0.900	0.497	4,933	0.005	0.105	0.950	0.438	4,903
0.111	0.135	0.799	0.692	4,964	0.022	0.093	0.948	0.575	4,919
0.117	0.148	0.761	0.687	4,990	0.001	0.090	0.952	0.536	4,932

Table 6.4 Comparison of the haplotype association estimation results between the prospective quadratic model with $n = 500$ for each control and case group)

			Prospective analysis				
Model	log OR(β_G)	ϕ	Bias	SE	95% coverage	Size/power	Of converges
Recessive	log(1.0)	–1.0	0.028	0.252	0.948	0.052	5,000
		–0.5	–0.021	0.264	0.948	0.052	5,000
		0.0	–0.016	0.276	0.957	0.043	5,000
		0.5	–0.038	0.304	0.950	0.050	5,000
		1.0	–0.043	0.328	0.952	0.048	5,000
	log(1.5)	–1.0	–0.009	0.241	0.947	0.375	5,000
		–0.5	0.012	0.250	0.951	0.398	5,000
		0.0	0.003	0.262	0.954	0.347	5,000
		0.5	0.010	0.282	0.953	0.331	5,000
		1.0	0.026	0.307	0.954	0.312	5,000
Dominant	log(1.0)	–1.0	0.010	0.131	0.947	0.053	5,000
		–0.5	–0.012	0.129	0.948	0.052	5,000
		0.0	–0.018	0.129	0.954	0.046	5,000
		0.5	–0.007	0.13	0.948	0.052	5,000
		1.0	0.005	0.132	0.948	0.052	5,000
	log(1.2)	–1.0	0.005	0.132	0.951	0.295	5,000
		–0.5	–0.012	0.130	0.949	0.252	5,000
		0.0	0.001	0.130	0.952	0.284	5,000
		0.5	0.012	0.131	0.956	0.310	5,000
		1.0	–0.008	0.133	0.953	0.256	5,000
Additive	log(1.0)	–1.0	0.011	0.102	0.947	0.053	5,000
		–0.5	–0.011	0.103	0.951	0.049	5,000
		0.0	–0.014	0.105	0.953	0.047	5,000
		0.5	–0.010	0.108	0.947	0.053	5,000
		1.0	–0.002	0.112	0.951	0.049	5,000
	log(1.2)	–1.0	0.011	0.101	0.951	0.475	5,000
		–0.5	–0.011	0.104	0.951	0.387	5,000
		0.0	–0.002	0.106	0.952	0.406	5,000
		0.5	0.015	0.108	0.952	0.450	5,000
		1.0	–0.008	0.110	0.953	0.339	5,000
Jointhaplotype	log(1.0)	–1.0	0.014	0.120	0.950	0.050	5,000
		–0.5	–0.007	0.120	0.952	0.048	5,000
		0.0	–0.020	0.120	0.952	0.048	5,000
		0.5	–0.006	0.120	0.959	0.041	5,000
		1.0	–0.004	0.122	0.951	0.049	5,000
	log(1.2)	–1.0	0.016	0.120	0.946	0.378	5,000
		–0.5	0.000	0.121	0.954	0.323	5,000
		0.0	0.000	0.121	0.948	0.326	5,000
		0.5	0.020	0.121	0.953	0.380	5,000
		1.0	–0.005	0.122	0.953	0.292	5,000

analysis with the retrospective analysis in case-control setting (genotype data generated under

Retrospective analysis									
Under HWE					Under HWD				
Bias	SE	95% coverage	Size/ power	Of converges	Bias	SE	95% coverage	Size/ power	Of converges
0.233	0.200	0.754	0.246	4,929	0.173	0.205	0.845	0.155	2,599
0.093	0.212	0.911	0.089	4,939	0.044	0.212	0.943	0.057	2,134
−0.039	0.218	0.958	0.042	4,905	−0.076	0.220	0.951	0.049	2,036
−0.234	0.238	0.851	0.149	4,914	−0.255	0.247	0.832	0.168	2,173
−0.454	0.255	0.580	0.420	4,893	−0.466	0.255	0.567	0.433	2,480
0.188	0.184	0.810	0.873	4,930	0.14	0.187	0.878	0.804	2,496
0.111	0.190	0.895	0.772	4,902	0.069	0.189	0.937	0.687	2,268
−0.024	0.194	0.955	0.514	4,919	−0.055	0.197	0.953	0.441	2,025
−0.194	0.201	0.872	0.198	4,901	−0.206	0.202	0.863	0.172	2,025
−0.384	0.212	0.584	0.048	4,895	−0.395	0.215	0.568	0.045	2,446
−0.076	0.121	0.900	0.100	4,930	−0.049	0.127	0.921	0.079	2,656
−0.056	0.121	0.928	0.072	4,937	−0.036	0.123	0.942	0.058	2,189
−0.017	0.12	0.955	0.045	4,907	0.000	0.121	0.957	0.043	1,997
0.051	0.121	0.933	0.067	4,910	0.058	0.121	0.925	0.075	2,046
0.121	0.123	0.829	0.171	4,873	0.121	0.123	0.827	0.173	2,310
−0.084	0.123	0.892	0.133	4,926	−0.055	0.126	0.927	0.173	2,779
−0.054	0.122	0.924	0.177	4,893	−0.031	0.123	0.942	0.231	2,257
−0.006	0.122	0.950	0.299	4,906	0.009	0.123	0.951	0.334	2,118
0.058	0.123	0.925	0.505	4,884	0.061	0.120	0.926	0.498	2,077
0.115	0.124	0.842	0.673	4,880	0.116	0.124	0.836	0.678	2,298
0.010	0.108	0.940	0.060	4,930	0.013	0.109	0.938	0.062	2,714
−0.011	0.106	0.947	0.053	4,937	−0.007	0.104	0.951	0.049	2,216
−0.015	0.104	0.953	0.047	4,906	−0.013	0.102	0.953	0.047	2,001
−0.010	0.103	0.953	0.047	4,910	−0.011	0.103	0.953	0.047	2,005
−0.002	0.102	0.959	0.041	4,885	−0.005	0.101	0.960	0.040	2,243
0.022	0.106	0.941	0.503	4,927	0.021	0.104	0.943	0.509	2,564
−0.006	0.106	0.946	0.409	4,900	−0.006	0.104	0.954	0.409	2,198
−0.002	0.105	0.951	0.413	4,913	−0.007	0.105	0.950	0.395	1,927
0.006	0.103	0.958	0.437	4,882	0.002	0.103	0.964	0.421	1,934
−0.025	0.099	0.956	0.306	4,899	−0.027	0.099	0.957	0.306	2,152
0.012	0.116	0.951	0.049	4,931	0.016	0.115	0.949	0.051	2,668
−0.011	0.115	0.951	0.049	4,939	−0.007	0.113	0.956	0.044	2,173
−0.021	0.116	0.952	0.048	4,900	−0.019	0.114	0.957	0.043	1,944
−0.011	0.116	0.953	0.047	4,908	−0.010	0.116	0.952	0.048	1,962
−0.006	0.118	0.949	0.051	4,884	−0.008	0.118	0.952	0.048	2,229
0.019	0.115	0.945	0.411	4,928	0.019	0.113	0.942	0.423	2,520
−0.003	0.116	0.949	0.337	4,895	−0.004	0.114	0.955	0.333	2,157
−0.003	0.116	0.951	0.337	4,916	−0.008	0.116	0.950	0.329	1,897
0.012	0.116	0.949	0.379	4,894	0.006	0.116	0.947	0.356	1,892
−0.012	0.116	0.951	0.302	4,902	−0.014	0.116	0.951	0.292	2,119

Table 6.5 Comparison of HLA allele prediction accuracy for the British 1958 birth cohort data (genotyped by Affymetrix 500K) between using the predictive models described in Leslie et al. (2008) and Li et al. (2011)

		Prediction accuracy (%) at high resolution in 4 digits (call rate %)			Prediction accuracy at intermediate resolution in 2 digits (call rate %)		
Gene	Predictive model	CT = 0	CT = 0.5	CT = 0.9	CT = 0	CT = 0.5	CT = 0.9
HLA-A	Leslie et al.	89	91 (93)	97 (58)	93	94 (94)	95 (29)
	Li et al.	96	96 (99)	98 (88)	97	97 (100)	99 (90)
HLA-B	Leslie et al.	82	85 (88)	93 (66)	84	87 (89)	94 (65)
	Li et al.	94	95 (98)	97 (82)	95	95 (100)	97 (89)
HLA_DRB1	Leslie et al.	72	76 (88)	83 (51)	86	90 (88)	95 (55)
	Li et al.	85	90 (81)	99 (38)	97	98 (99)	98 (77)
HLA_DQB1	Leslie et al.	77	80 (88)	93 (29)	90	91 (89)	97 (31)
	Li et al.	94	95 (98)	98 (80)	99	99 (100)	99 (97)

rare HLA alleles. However, since HLA alleles are found on multiple haplotype backgrounds (Walsh et al. 2003), they can be tagged or predicted by SNP haplotypes. Leslie et al. have developed an IBD-based method to predict HLA alleles using *phased* haplotypes from parent-offspring trio data (Leslie et al. 2008). The method was applied to build models predicting HLA-A, -B, -C, -DRB1, and -DQB1 alleles in high and intermediate resolution using the data of 45 parent-offspring trios with European ancestry from Utah. The models were then validated using the British 1958 birth cohort data. The prediction accuracy ranged from 72 to 89% for predicting high-resolution alleles and 84 to 93% for predicting intermediate resolution HLA alleles. The requirement of phased haplotypes limits the available data to use and thereby limits the accuracy to predict rare HLA alleles for the IBD-based method.

To overcome this limitation, we introduced a new likelihood-based method that builds predictive models using genotypes from independent samples (Li et al. 2011). The key idea was similar to our haplotype inference method as described in Sect. 6.2. Rather than constructing haplotypes of SNPs alone, here we constructed haplotypes of SNPs together with HLA alleles. Then predictive models for HLA alleles from SNP genotypes were built based on the constructed haplotypes. The predictive models were validated using an independent data set. The prediction accuracy was measured by the percentage of correctly predicted HLA alleles in the validation set. Applying the method to the data generated by a GWAS of hematopoietic cell transplant (HCT) outcomes of cohort of ~1,500 patient-donor pairs, we used the unrelated donor data (about a half of the cohort donors) to build the predictive models and used the related donor data to validate the prediction. The prediction accuracy ranged from 79 to 97% for predicting HLA alleles in high resolution and 93–98% for predicting HLA alleles in intermediate resolution. To compare our results with the results from (Leslie et al. 2008), we applied the protective models to the British 1985 birth cohort. Our prediction accuracy was 10% higher than the prediction accuracy using the method of (Leslie et al. 2008) on average (Table 6.5). The prediction accuracy generally depends on the copies of specific alleles in the training data set used for building the predictive model. Figure 6.1 shows the relationship between the

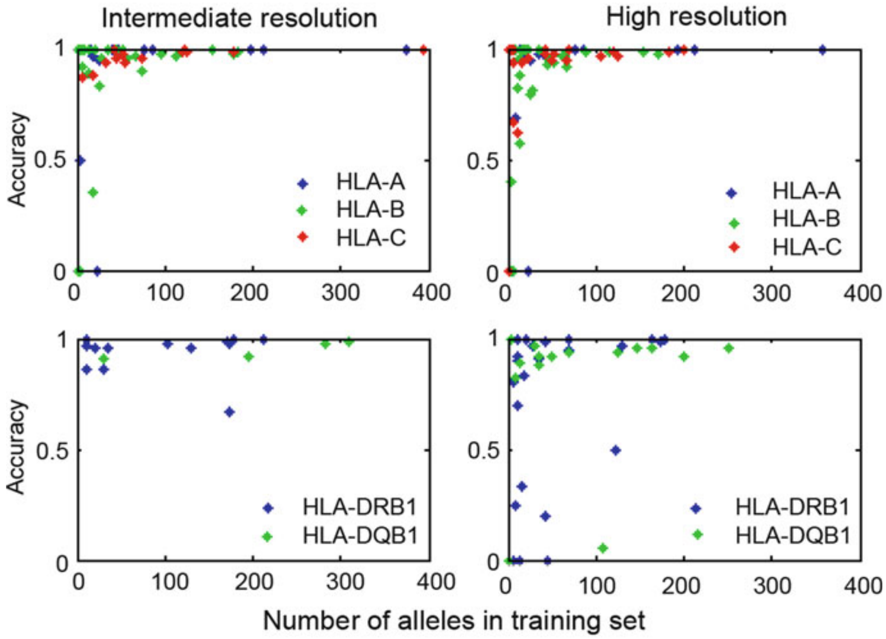

Fig. 6.1 Relationship between the prediction accuracy for each allele in the validation set and the number of copies of the allele observed in the training set

predictive accuracy and the number of observed alleles in the training set used to build the predictive model.

To evaluate the performance of these predictive models further, we (Zhang et al. 2011) investigated several practical issue and concluded that (1) including imputed SNPs in MHC using HapMap data in addition to genotyped ones can improve the accuracy of prediction; (2) SNPs generated by different platforms yield a comparable accuracy of predictions; (3) the prediction models built based on the data from a particular population can be used for predicting HLA alleles from SNP genotypes generated from different study populations of the same ethnicity; and (4) combining all available SNP and HLA data that observed from multiethnic populations to build prediction models generally improves the prediction accuracy as well.

Consider a random sample of n individuals. On each individual, the genotype of an HLA gene is denoted by $h_i = \dot{h}_i\ddot{h}_i$, where $\dot{h}_i$ and $\ddot{h}_i$ are two HLA alleles of the i-th individual. Suppose that on each individual sample, we have q SNPs flanking the HLA locus, denoted as $G_i = (g_{i1}, g_{i2}, \ldots, g_{iq})$. In this section, we denote the pair of haplotypes of G_i as $(\dot{H}_i, \ddot{H}_i)$, where $\dot{H}_i$ and $\dot{h}_i$ are on the same chromosome and $\ddot{H}_i$ and $\ddot{h}_i$ are on the other chromosome. Given the expected LD between the HLA gene and flanking SNPs, the joint distribution of the HLA and SNP genotypes is expressed as

$$f(h_i, G_i) = \sum_{(\dot{h}_i\dot{H}_i, \ddot{h}_i\ddot{H}_i) \in S(h_i, G_i)} f(\dot{h}_i\dot{H}_i, \ddot{h}_i\ddot{H}_i) = \sum_{(\dot{h}_i\dot{H}_i, \ddot{h}_i\ddot{H}_i) \in S(h_i, G_i)} f(\dot{h}_i\dot{H}_i) f(\ddot{h}_i\ddot{H}_i), \qquad (6.11)$$

where $S(h_i, G_i)$ is the set of all haplotype pairs that give arise to the genotypes (h_i, G_i), $f(\dot{h}_i\dot{H}_i)$ and $f(\ddot{h}_i\ddot{H}_i)$ are the distributions of joint haplotype of HLA and SNPs, and the last equation holds under the Hardy-Weinberg equilibrium (HWE). The haplotype $\dot{h}_i\dot{H}_i$ is assumed to have a multinomial distribution expressed as

$$f(\dot{h}_i\dot{H}_i) = \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}_i\dot{H}_i = \underline{hH})}, \tag{6.12}$$

where $\Pr(\underline{hH})$ is the frequency of observing a joint haplotype of HLA and SNPs, $\underline{hH}$, the indicator function $I(\dot{h}_i\dot{H}_i = \underline{hH})$ equals to one if the inside equality holds and zero otherwise, and Θ is the set of all possible joint haplotypes of HLA and SNPs. The similar formulation is for the distribution of $\ddot{h}_i\ddot{H}_i$. To estimate the haplotype frequency $\Pr(\underline{hH})$, we employed the likelihood method, via maximizing the following log-likelihood function:

$$l = \sum_{i=1}^{n} \log\left(\sum_{(\dot{h}_i\dot{H}_i, \ddot{h}_i\ddot{H}_i)\in S(h_i, G_i)} \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}_i\dot{H}_i = \underline{hH}) + I(\ddot{h}_i\ddot{H}_i = \underline{hH})} \right). \tag{6.13}$$

The detailed estimation procedure was described in (Li et al. 2003).

According to the Bayesian rule, the predictive probability for HLA alleles given SNP genotypes can be written as

$$\Pr(\dot{h}\ddot{h} \mid G) = \frac{f(\dot{h}\ddot{h}, G)}{\sum_{\dot{h}\ddot{h}} f(\dot{h}\ddot{h}, G)} = \frac{\sum_{(\dot{h}\dot{H}, \ddot{h}\ddot{H})\in S(h,G)} \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}\dot{H} = \underline{hH}) + I(\ddot{h}\ddot{H} = \underline{hH})}}{\sum_{\dot{h}\ddot{h}} \sum_{(\dot{h}\dot{H}, \ddot{h}\ddot{H})\in S(h,G)} \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}\dot{H} = \underline{hH}) + I(\ddot{h}\ddot{H} = \underline{hH})}}, \tag{6.14}$$

where the first summation in the denominator is over all possible genotypes at the HLA locus and the second summation in the denominator and the first summation in the numerator are over all possible joint haplotype pairs that are consistent with the genotypes of HLA and SNPs.

The above predictive probability for any pair of HLA alleles takes values between zero and one. It is possible for more than one pair of HLA alleles to be associated with a positive predictive probability. In practice, the pair associated with the highest predictive probability is then called the predicted HLA result. If we accept all predicted results regardless of their associated predictive probabilities, then we make calls for all samples and the call rate is 100%. In the case of requiring higher confidence on prediction results, one could make a call only if the predicted result is associated with a probability above a threshold, e.g., 0.5 or 0.9. In this case, the call rate is possibly less than 100%.

An important step in building a successful predictive model is to select a minimum number of SNPs that predict HLA alleles with the most accuracy. To achieve this goal, we constructed an objective function based upon the Akaike information criterion (AIC) (Koehler and Murphree 1988), which maximizes the log-likelihood function and penalizes on the number of additional haplotype parameters to be estimated,

$$Q = -\sum_{i} \log\left(\frac{\displaystyle\sum_{(\dot{h}_i\dot{H}_i,\ddot{h}_i\ddot{H}_i)\in S(h_i,G_i)} \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}_i\dot{H}_i=\underline{hH})+I(\ddot{h}_i\ddot{H}_i=\underline{hH})}}{\displaystyle\sum_{\dot{h}\ddot{h}} \sum_{(\dot{h}\dot{G}_i,\ddot{h}\ddot{G}_i)\in S(h,G_i)} \prod_{\underline{hH}\in\Theta} \Pr(\underline{hH})^{I(\dot{h}_i\dot{H}_i=\underline{hH})+I(\ddot{h}_i\ddot{H}_i=\underline{hH})}} \right) + (m-k), \quad (6.15)$$

where the first logarithmic term in the above equation is the negative log likelihood of predictive probabilities given all SNP genotypes in the training set and the second term equals the difference of the number of haplotypes of HLA-SNP (m) and the number of HLA alleles (k). Note that the first term of the objective function decreases when the number of SNPs in the model increases. Increasing the number of SNPs in the model increases the number of haplotypes and, therefore, increases the number of parameters for estimation. To avoid over fitting the model, the second term is added in the objective function. Our goal is to find the optimal set of SNPs to minimize the objective function.

In the current chip design, there are none or a few SNP probes within the HLA loci. Thus, we had to include the SNPs from flanking region of each HLA locus. Linkage disequilibrium between SNPs and HLA alleles generally decreases with increasing distance between the SNP and HLA locus. Choosing the boundary for SNPs to be used in predictive model was guided by evaluating the objective function. The boundary was set when the objective function reached to the minimum. To make the selection process efficient, we proposed an SNP selection procedure starting from the SNPs within HLA locus and gradually expending to the SNPs at both sides of flanking region by adding one SNP at a time using a combination of forward selection and backward elimination scheme.

6.5 Summary

Haplotype-based analyses are useful in genetic associations. While they are unlikely to be applicable for all circumstances, they have their own utilities. In this chapter, we have identified several areas where more complex haplotype analysis, debatably more sophisticated genetic analyses, is quite helpful. To reiterate, haplotype analysis is desirable if the focus is on a well-defined biological unit such as multiple SNPs within a specific gene, specific exon, or regulatory region (introns, or promoters, or even intergenetic regions) and the goal is to identify functional elements locating on the same chromosome. Discoveries of any haplotype-based associations are more readily to be validated with functional studies or sequence experiments. At risk of being criticized, the haplotype analysis is more like "surgeon's knife," which is applicable only for specific situations.

6.6 Software

HPlus (http://qge.fhcrc.org/hplus/) performs the haplotype analyses that include estimating haplotype frequency, inferring individuals' haplotypes from SNP genotypes, and correlating SNP haplotypes with a phenotype that is either quantitative, or binary, or time-to-event.

MAGprediction (http://qge.fhcrc.org/MAGprediction/) predicts highly polymorphic gene alleles, HLA alleles in particular, using unphased SNP data. The software includes two models: a training module that builds prediction models using user-provided HLA and SNP data and a prediction module that predicts the HLA alleles based on either the several available models built by us or on results from the training module.

References

Altshuler D, Brooks LD, Chakravarti A, et al. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.

Asano K, Matsushita T, Umeno J, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat Genet. 2009;41(12): 1325–9. doi:10.1038/ng.482.

Bentley DR, Deloukas P, Dunham A, et al. The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. Nature. 2001;409(6822):942–3.

Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12(10):703–14. doi:10.1038/nrg3054.

Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends Genet. 2003;19(3):135–40.

Chen J, Peters U, Poster C, et al. A haplotype-based test of association using data from cohort and nested case-control epidemiologic studies. Hum Hered. 2004;58:18–29.

Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol. 1990;7:111–22.

Collins A, Morton NE. Mapping a disease locus by allelic association. Proc Natl Acad Sci U S A. 1998;95(4):1741–5.

Daly MJ, Rioux JD, Schaffner SF, et al. High-resolution haplotype structure in the human genome. Nat Genet. 2001;29:229–32.

Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet. 2003;73:1316–29.

Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995;12(5):921–7.

Fellay J, Frahm N, Shianna KV, et al. Host genetic determinants of T cell responses to the MRKAd5 HIV-1 gag/pol/nef vaccine in the step trial. J Infect Dis. 2011;203(6):773–9. doi:10.1093/infdis/jiq125.

Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225–9.

Goldstein DB. Islands of linkage disequilibrium. Nat Genet. 2001;29:109–11.

Green ED, Cox DR, Myers RM. The human genome project and its impact on the study of human disease. In: Vogelstein B, Kinzler KW, editors. The genetic basis of human cancer. New York: McGraw-Hill, Health professional division; 1998. p. 33–63.

Gyapay G, Morissette J, Vignal A, et al. The 1993-94 Genethon human genetic linkage map. Nat Genet. 1994;7:246–339.

Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered. 1995;86:409–11.

Hirschfield GM, Liu X, Xu C, et al. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. N Engl J Med. 2009;360(24):2544–55. doi:10.1056/NEJMoa0810440.

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529. doi:10.1371/journal.pgen.1000529.

King RA, Rotter JI, Motulsky AG. The genetic basis of common diseases, Oxford monographs on medical genetics, vol. 44. 2nd ed. Oxford/New York: Oxford University Press; 2002.

Koehler AB, Murphree ES. A comparison of the Akaike and Schwarz criteria for selecting model order. Appl Stat. 1988;37(2):187–95.

Lander ES, Botstein D. Mapping complex genetic traits in humans: new method using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol. 1986;51:49.

Larsen CE, Alper CA. The genetics of HLA-associated disease. Curr Opin Immunol. 2004;16(5):660–7. doi:10.1016/j.coi.2004.07.014.

Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008;82(1):48–56.

Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–33.

Li S, Khalid N, Carlson C, et al. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. Biostatistics. 2003;4(4):513–22.

Li SS, Cheng JJ, Zhao LP. Empirical vs Bayesian approach for estimating haplotypes from genotypes of unrelated individuals. BMC Genet. 2007;8:2.

Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34. doi:10.1002/gepi.20533.

Li SS, Wang H, Smith A, et al. Predicting multiallelic genes using unphased and flanking single nucleotide polymorphisms. Genet Epidemiol. 2011;35(2):85–92. doi:10.1002/gepi.20549.

Lin DY. Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol. 2004;26:255–64.

Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc. 2006;101:89–104.

Lin S, Cutler DJ, Zwick ME, et al. Haplotype inference in random population samples. Am J Hum Genet. 2002;71(5):1129–37.

Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet. 1995;56:799–810.

Murray JC, Buetow KH, Weber JL, et al. A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). Science. 1994;265(5181):2049–54.

NIH/CEPH Collaborative Mapping Group. A comprehensive genetic linkage map of the human genome. [Review]. Science. 1992;258(5079):67–86.

Niu T, Qin ZS, Xu X, et al. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet. 2002;70(1):157–69.

Patil N, Berno AJ, Hinds DA, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001;294(5547):1719–23.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979;66(3):403–11.

Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics. 1991;47(3):825–39.

Qin ZS, Niu T, Liu JS. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet. 2002;71:1242–7.

Reveille JD, Sims AM, Danoy P, et al. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet. 2010;42(2):123–7. doi:10.1038/ng.513.

Sabeti PC, Varilly P, Fry B, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–18. doi:10.1038/nature06250.

Satten GA, Epstein MP. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol. 2004;27:192–201.

Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol. 2004;27:348–64.

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44. doi:10.1086/502802.

Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol. 2005;29:108–27.

Stefansson H, Ophoff RA, Steinberg S, et al. Common variants conferring risk of schizophrenia. Nature. 2009;460(7256):744–7. doi:10.1038/nature08186.

Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 2003;73(5):1162–9.

Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001;68(4):978–89.

Stram DO, Pearce CL, Bretsky P, et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered. 2003;55:179–90.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78. doi:10.1038/nature05911.

Tse KP, Su WH, Chang KP, et al. Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am J Hum Genet. 2009;85(2):194–203. doi:10.1016/j.ajhg.2009.07.007.

Tzeng J-Y, Wang C-H, Kao J-T, et al. Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet. 2006;78:231–42.

Valle T, Tuomilehto J, Bergman RN, et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care. 1998;21(6): 949–58.

Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science. 2001; 291(5507):1304–51.

Walsh EC, Mather KA, Schaffner SF, et al. An integrated haplotype map of the human major histocompatibility complex. Am J Hum Genet. 2003;73(3):580–90.

Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998;280:1077–82.

Wijsman EM. A deductive method of haplotype analysis in pedigrees. Am J Hum Genet. 1987;41:356–73.

Xiong MM, Guo SW. Fine-scale genetic mapping based on linkage disequilibrium-theory and applications. Am J Hum Genet. 1997;60(6):1513–31.

Zeng D, Lin DY, Avery CL, et al. Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics. 2006;7(3):486–502. doi:10.1093/biostatistics/kxj021.

Zhang XC, Li SS, Wang H, et al. Empirical evaluations of analytical issues arising from predicting HLA alleles using multiple SNPs. BMC Genet. 2011;12:39. doi:10.1186/1471-2156-12-39.

Zhao LP, Aragaki C, Hsu L, et al. Mapping complex traits with single nucleotide polymorphisms. Am J Hum Genet. 1998;63:225–40.

Zhao LP, Li SS, Khalid N. A method for assessing disease associations with SNP haplotypes and environmental variables in case-control studies. Am J Hum Genet. 2003;72:1231–50.

Chapter 7
Analytical Approaches for Exome Sequence Data

Andrew Collins

Abstract Sequencing the 1% of the genome coding for proteins (the exome) offers a powerful and often cost-effective route to identifying genetic mutations underlying Mendelian disease. It is possible that exome sequencing in a relatively small number of individuals showing ‘extreme’ phenotypes or more familial subtypes of complex disease may also be productive. Larger-scale exome and whole genome sequencing studies offer the potential to interrogate the cumulative impact of the numerous rare variants presumed to underlie a substantial proportion of complex disease susceptibility. Exome and, particularly, whole genome sequencing studies yield enormous amounts of data and pose many analytical challenges. Aside from issues concerning the production of high-quality sequence reads and the management and manipulation of huge databases, a major concern, in the early stages of analysis, is the reliable alignment of the short sequence reads against a reference genome. A wide range of algorithms and software tools for alignment have been developed and implemented for this most critical step in every analysis ‘pipeline’. A similarly rich set of platforms and analytical tools are available to facilitate the reliable calling of DNA variants. Given the excellent resources now available, the production of a well-characterised database cataloguing novel and known variants in an individual exome is achievable. However, the difficulty of teasing out causal variants from the vast amount of neutral or irrelevant variation presents the greatest challenge. I review here the techniques and tools that have been developed and applied for the analysis of exome data. Exome mapping of genes involved in Mendelian disease has met with considerable success thus far, while applications to complex traits look promising given analysis of sufficiently large numbers of case and control exomes.

A. Collins (✉)
Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine,
University of Southampton, Duthie Building (808), Tremona Road,
Southampton, SO16 6YD, UK
e-mail: arc@soton.ac.uk

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_7, © Springer Science+Business Media Dordrecht 2012

Keywords Complex disease • Exome sequencing • Mendelian disease • Sequence alignment • Variant annotation

7.1 Introduction

Thousands of genetic variants for both Mendelian diseases and complex traits have been identified as causal or associated with disease phenotypes in recent years. These have usually been identified through linkage mapping, in the case of Mendelian disease, and candidate gene studies or genome-wide association studies (GWAS) in the case of complex traits. For complex diseases the majority of the implicated single nucleotide polymorphism (SNP) variants are associated indirectly with disease, usually to a genomic region. Because these regions can be large and/or inter-genic, GWAS associations may or may not indicate whether a specific gene is compromised and involved in disease. In contrast, sequencing enables the identification of all variants in a genome or genomic region such that an individual variant can, in favourable circumstances, be firmly identified as causal. For this reason exome sequencing and whole genome sequencing are already revolutionising the way genetic studies are undertaken.

Recent years have seen dramatic changes in the development and application of DNA sequencing technology. The traditional Sanger sequencing method employing capillary electrophoresis remains the 'gold standard' in terms of the length of the reads and the accuracy of the sequence (Harismendy et al. 2009). However, 'next-generation sequencing' (NGS) methods generate 3 or 4 orders of magnitude more sequence at greatly reduced cost compared to the Sanger approach. These methods sequence DNA molecules spatially separated in flow cell and attached to a solid surface. The process employs optical imaging to record the sequential addition of nucleotides in the sequencing reaction. This enables millions of sequencing reactions to take place in parallel. The first massively parallel NGS platform was launched in 2005 (Majewski et al. 2011). NGS radically overcomes the problem of limited scalability of the Sanger approach (Reis-Filho 2009; Lander 2011) and is capable of generating hundreds of mega- to giga-base pairs (bp) of nucleotide sequence in a single run. Millions of overlapping sequence reads are then aligned and compared to a reference genome to identify differences (polymorphisms). Targeted sequencing of genomic regions of particular interest, of which the most important is undoubtedly the entire exome (the protein-coding exons of all genes), has benefits with respect to reduced cost, data management and increased sequence coverage (for a given quantity of DNA). Exome sequencing typically involves sequencing the ends of fragments from the sheared sample DNA – either one end (single-end sequencing) or both ends (paired-end sequencing) of the fragments. The sequence read lengths are typically in the range of 35–150 bp for Illumina platforms (http://www.illumina.com/applications/sequencing/targeted_resequencing.ilmn) and ~400 base pairs for the Roche 454 sequencer (http://www.roche.com/products/product-list.htm?type=researchers&id=4). The exome comprises only ~1% of the genome (~30 Mb), so an average 'depth' of coverage of the exome of 75 can be achieved with 3 Gbp of sequence, whereas 90 Gbp would be required for 30-fold-depth coverage of the whole genome (Majewski et al. 2011; Bainbridge et al. 2010).

The exome is the best understood component of the genome for relating sequence to function and, similarly, to directly link genetic variants with disease causality (Kumar et al. 2011). For Mendelian disorders exome sequencing offers a powerful route to identifying the underlying allelic variants since the majority of this class of disease genes are known to disrupt protein-coding sequences. Kryukov et al. (2007) have shown that most rare non-synonymous (missense) alleles are likely to be deleterious, unlike the majority of noncoding sequences. The exome is therefore particularly enriched for variants underlying Mendelian traits. There is also increasing evidence that exome sequencing offers a route to understanding complex disease. For example, it has been shown that rare variants are over-represented in genes already identified (usually by GWAS) as containing common variants involved in complex disease. Johansen et al. (2010) determined a significant burden ('mutation skew') of 154 rare missense or nonsense variants in 438 individuals with hypertriglyceridemia, compared to a significantly lower burden in controls, within four genes known to contain common variants for this condition. Support for the observation of rarer alleles with potentially higher disease penetrance residing within genes implicated by GWAS comes from the study by Rivas et al. (2011). Working on the inflammatory bowel disease (IBD) phenotypes, the authors identified novel rare variants which contribute a greater component to the population risk variance than the known common IBD variants in the *CARD9*, *NOD2*, *CUL2* and *IL18RAP* genes. Lehne et al. (2011) questioned whether missing the regulatory elements that may impact disease phenotype, but are situated outside the exome sequence regions, would reduce the value of applying exome sequencing to complex disease. For most of complex diseases examined, the authors found that most of the association signal from 'suggestive' common variants was found within the coding regions rather than introns. Although they did not consider rare variation directly, the work supports exome sequencing as a strategy to search for genetic variation associated with complex disease.

Despite its evident advantages and early successes, exome sequencing has a number of disadvantages and problems, aside from the obvious lack of information from the bulk of the noncoding genome. Exon capture requires the use of complementary nucleic acid 'baits' to trawl sequence reads from specific exons. Since these are 'small' targets, this can result in uneven coverage of exonic regions, and the baits themselves are only as complete as the information derived from gene annotation and other reference databases. There is also a degree of low-depth hybridisation away from the targets in non-exonic regions although the overlap of sequence reads extending a short distance either side of the bait probes provides some information on adjacent regions. There is a trend towards increasing the coverage of exonic and adjacent regions in the newer products. Perhaps more important than concerns about coverage are a wide range of data analytical considerations, reviewed here.

7.2 Strategies for Exome Projects

The strategy chosen for an exome sequencing study depends on the known, expected or hypothesised genetic mode of inheritance. The costs and analytical challenges of sequencing hundreds of exomes to pursue the complete spectrum of rare variation underlying complex disease are likely to be prohibitive for all but large consortia for the foreseeable future. At the other end of the spectrum, highly successful studies focussed on a small number of related individuals have been achieved for Mendelian diseases. Between these two extremes is perhaps the most intriguing prospect: sequencing a small number of affected relatives showing relatively strong familial patterns for a complex trait and/or focussing on a distinct disease subtype or individuals showing an 'extreme' phenotype of a common disease might identify important rare variation. Success depends on the existence of forms of complex disease closer to the Mendelian end of the disease spectrum, and strategies include focus on individuals with particularly severe forms of a disease and/or markedly early onset. For complex diseases there remains a substantial degree of uncertainty about how best to design such studies, but I consider here some of the findings to date.

7.2.1 Mendelian Disorders

Fewer than half of the allelic variants underlying monogenic diseases showing a Mendelian pattern of inheritance have been identified. The difficulty with finding many of these genes arises from the rarity of affected cases or case families, the existence of similar phenotypes determined by independent mutations (locus heterogeneity) and the reduced reproductive fitness limiting the further analysis of key pedigrees. Many of these more difficult diseases arise as *de novo* mutations and are not therefore amenable to linkage analysis. However, exome sequencing offers a route to progress and initial applications, focussed on a number of Mendelian disorders, have identified high-penetrance genes through sequencing a very small number of affected family members. Ng et al. (2009) were the first to demonstrate the utility of exome sequencing to identify Mendelian disease variants. As proof of principle, the authors sequenced the exomes of four unrelated cases with Freeman-Sheldon syndrome, a disease for which the causal variant was known, and eight control samples. The authors filtered out common and presumed unimportant variation identified in HapMap and dbSNP and demonstrated that disease variants could be mapped solely by exome sequencing of a few cases. The gene for Miller syndrome (Ng et al. 2010a) was the first example of a gene found for a disease of unknown cause. The DHODH gene was mapped using four affected cases in three independent pedigrees, data filtered against public SNP variant databases, and verified by Sanger sequencing in three additional Miller families. To maximise the chance of identifying the gene, the authors considered a dominant model with at least one novel non-synonymous SNP, splice variant or coding indel. Their recessive model required genes with at least two novel variants which were either in the same position (homozygous) or in different positions (as a possible compound heterozygote but conditional on, unknown, phase). The success of this enterprise depended to a large extent on the choice of disease. Miller syndrome is a very rare Mendelian disease, and so causal variants were unlikely to be present in reference databases or control exomes. Mapping a rare recessive gene is easier than a dominant gene because fewer genes within the affected individual's exome will have two novel or rare non-synonymous variants. The lack of genetic heterogeneity in the sample of individuals studied was also advantageous, and the authors emphasise the importance of ethnic uniformity in the ancestry of affected cases (Europeans in this case) reducing the likelihood of genetic heterogeneity.

Strategies that might accelerate the mapping of Mendelian disorders in the future include, for recessive models, identifying genes within shared tracts of homozygosity to reduce the pool of potential candidate variants for further consideration. Krawitz et al. (2010) introduced identify-by-descent filtering to map the recessive gene for hyperphosphatasia mental retardation syndrome (HPMRS or Mabry syndrome) in a family with three affected siblings. They developed a hidden Markov model to identify regions with shared identical, maternal and paternal haplotypes but not necessarily derived from a common ancestor. They were then able to identify whether each sibling had the same (identity by descent = 2) homozygous or heterozygous genotype. This process reduced the pool of candidate genes with mutations in all three sibs from 14 to 2 and led to the identification of the PIGV gene as causal.

7.2.2 De Novo Variants

For 'sporadic' disease sequencing of unaffected parents may facilitate rapid identification of important *de novo* mutations involved in disease. Girard et al. (2011) sequenced exomes and parents of 14 schizophrenia probands with no previous family history and identified 15 *de novo* mutations in eight probands. This is a higher *de novo* mutational burden than the 'background' mutation rate as indicated by the 1000 Genomes Project. Four of the 15 mutations were predicted to lead to a premature stop codon in genes hypothesised to have a role in the disease.

7.2.3 Cancer Germline and Tumour Studies

A route to further understand the genetic basis of cancer is offered by the exome sequencing in both germline and tumour DNA from the same patient and searching (by subtraction of the germline variants) for novel somatic mutations. An early success for this approach is described by Tiacci et al. (2011) who exome-sequenced germline and tumour DNA from an index patient with hairy-cell leukaemia (HCL). The findings included a somatic heterozygous mutation in the BRAF gene which was known to produce an oncogenic protein. Remarkably, the same variant was identified by Sanger sequencing as present in all 47 additional HCL patients they were screening but in none of their 195 patients with other forms of peripheral B-cell lymphoma or leukaemia. The power of this approach to identify recurrent somatic mutations driving further downstream somatic changes was clearly demonstrated. The findings also support BRAF mutation screening as a diagnostic tool to distinguish HCL from other B-cell lymphomas and identify HCL as a clinically distinct entity from other ‘HCL-like’ disorders.

7.2.4 Rare Variants in Families: Extreme Phenotypes

Feng et al. (2011) consider strategies for mapping rare variants in complex disease in the context of family data. The authors recognise the critical issues which reduce power, namely, locus heterogeneity (McClellan and King 2010), allelic heterogeneity (2,000 pathogenic mutations have been reported in BRCA2), problem of phenocopies (affected individuals in a family that do not share the predisposing mutations) and apparent oligogenic patterns of inheritance due to segregation of many common moderate-risk loci. Nevertheless, Cirulli and Goldstein (2010) argue that family-based designs, particularly for families showing phenotypes from the extremes of a trait distribution, are most likely to achieve success for complex traits until the costs of sequencing reduce sufficiently to favour very large case-control designs. Simulations support a two-stage design with sequencing of two affected individuals per pedigree that are not too closely related to generate an excessive number of false-positive genes or too distantly related to increase the risk of including a phenocopy in the comparison.

7.2.5 Rare Variants in Large Cohorts: Mutational Load

Cooper and Shendure (2011) consider the interpretive challenge of the ‘multiple hypothesis testing’ problem presented by the enormous number of variants identified in genome sequences and the abundance of false discoveries. They argue that experimental or computational approaches to assess variant function can provide estimates of the prior probability that a given variant is phenotypically important, thereby boosting discovery power. Such empowering classifiers include SIFT scores that use ‘evolution as the best measure of deleteriousness’, the observation that sequences not removed by natural selection are likely to be important. Application of a comprehensive range of functional and predictive tools is likely to be required for complete characterisation of important low-frequency variation identified in large cohorts of patients with common forms of disease. Evolutionary models predict that rare deleterious mutations spread across a large number of genes may have a cumulative effect (mutational load) to increase susceptibility to complex disease.

In this scenario a given mutation may be present in only a few individuals and have a negligible effect on trait variation, but, in combination with many similar variants, the burden of mutation may underlie causality (Howrigan et al. 2011). Pooled association tests and collapsing methods (Price et al. 2010; Dering et al. 2011) provide routes to testing mutational burden in large-scale genetic studies.

7.3 Exome Data

Data from a sequencer are typically presented in FASTQ format in which there are four lines per read comprising sequence identification labels, raw sequence and quality scores for each of the bases in the sequence (http://en.wikipedia.org/wiki/FASTQ_format). The quality score represents, as a single ASCII character, the probability (p) that the base call it refers to is incorrect. The Sanger version of the Phred quality score is $Q_{\text{sanger}} = -10 \log_{10} p$. Two such FASTQ files are generated for paired-end sequencing with sequential entries corresponding to the sequenced ends of each DNA fragment. Li et al. (2009a) describe the now standard 'sequence alignment/map' (SAM) format for storing short read alignments and mapping coordinates against a reference sequence. A software package (SAMtools) is used for processing such files and has options for positional sorting, indexing, format conversion and calling and viewing variants. The standardised format allows for efficient capture of read and alignment information by defining codes that characterise aligned sequences and identified variations from the reference sequence. These include, for example, codes to represent matches and mismatches, insertions, deletions and sequences with 'soft' and 'hard' clipping to represent non-matched sequences which are either present or missing from the alignment. Their CIGAR format provides a compact way of storing good alignments and also representing bases misaligned to the reference genome. The SAM format has a binary equivalent file (BAM file) which improves processing performance by supporting more rapid retrieval of aligned sequences in specific genomic regions.

7.3.1 Sequence Alignment

Accurate alignment of short read sequences against a reference genome is the most critical step towards cataloguing the polymorphisms represented in a sample. The process requires a reliable reference genome with known sequence and millions of short reads from the sample genome. Many algorithms have been developed to align sequence reads against the reference genome. Li and Homer (2010) and Ruffalo et al. (2011) survey the range of sequence alignment packages. Short read alignment packages include Bowtie (Langmead et al. 2009), BWA (Li and Durbin 2009), MAQ (Li et al. 2008), mrsFAST (Alkan et al. 2009), Novoalign (http://www.novocraft.com/main/index.php), SHRiMP (Rumble et al. 2009) and SOAPv2 (Li et al.

2009b). Of these, BWA is one of the most frequently used aligners. It exploits indexing built using the Burrows-Wheeler transformation (Burrows and Wheeler 1994) which enables fast searching and generates a quality score that can be used to reject poorly supported alignments. Ruffalo et al. undertook a simulation-based comparison and noted that the different approaches trade off speed and accuracy to optimise detection of different variant classes. Some algorithms were more efficient at different stages in the alignment process. For example, BWA and SOAP were found to align genomes quickly but required significant time to index the genome, whereas Novoalign required less time for indexing time but performance showed greater dependence on the number of reads. Novoalign offers high sensitivity and specificity with respect to accuracy of alignments and uses information on base qualities at all stages in the alignment (Li and Homer 2010) although this impacts on speed of the alignment. However, higher performance can be achieved by running the message passing interface (MPI) version on a computer cluster and exploiting multithreading.

7.3.2 Variant Calling

Given an aligned set of reads, it is essential to identify and 'mark' duplicate reads so that they do not influence variant calling. Tools to achieve this include PICARD (http://sourceforge.net/apps/mediawiki/picard/index.php?title=Main_Page) and SEAL (Pireddu et al. 2011), an alignment tool which combines BWA with the detection and removal of duplicate reads. Duplicates are likely to be PCR artefacts from the library preparation stage or optical duplicates from the sequencer. Duplicates are most simply defined as those reads that map to exactly the same locations. Other quality control preprocessing includes base quality score recalibration (applied to a BAM file) (http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration). This procedure recalibrates the scores to more accurately reflect the probability of mismatching the reference genome. The Genome Analysis Tool Kit (GATK) provides quality score recalibration which targets not only overall base quality inaccuracy but identifies higher quality subsets of bases by accounting for decline in base quality known to occur towards the ends of sequence reads.

Tools such as GATK and SAMtools are capable of identifying short indels in exome data, but accurate characterisation of indels in exome data is challenging. For example, short indels tend to occur in the vicinity of tandem repeats, but accurate alignment in these regions is difficult. Furthermore, where an indel is present, it may create local misalignments against the reference sequence which can generate false SNP calls. Therefore, local realignment around indels is required to minimise the number of mismatching bases (http://www.broadinstitute.org/gsa/gatkdocs/release/org_broadinstitute_sting_gatk_walkers_indels_IndelRealigner.html). Local realignment aims to resolve regions with misalignments caused by indels into clean reads, prior to applying tools to identify the variant content of the exome. Calling variants while using the information from more than one exome simultaneously increases the quality of variant calls. GATK's UnifiedGenotyper module employs a Bayesian genotype likelihood model to derive the most likely genotypes as applied to multiple samples simultaneously. The program also generates a posterior probability for a segregating variant allele as well as genotype at each locus.

VarScan (Koboldt et al. 2009, http://varscan.sourceforge.net/) is designed for identifying SNPs and indels in NGS data and is particularly suited to filtering in tumour-normal (tumour-germline) paired samples. Given such paired data, VarScan tests the somatic status of each variant and classifies them as germline, somatic or loss of heterozygosity by comparing the read counts between samples. VarScan uses the 'pileup' files of variant output from the SAMtools program from the germline and tumour DNAs simultaneously. Variant positions shared between both files meeting the minimum read depth coverage are compared and variants classified accordingly. Filtering against a germline sample of variants has obvious benefits in terms of reducing variant volume and complexity in the expectation of identifying recurrent 'driver' mutations that underlie the disease.

7.3.3 Filtering and Identifying Disease Susceptibility Genes

Sets of variant calls from an exome sequence include a large number of false positives. Suggested quality control filters, as implemented, for example, in the GATK program, include removal of variants at sites with low mapping quality scores and removal of apparent heterozygotes in which one allele is supported by less than 30% of sequence reads, variants not supported by reads mapping to both strands (strand bias). A significant difference of NGS from traditional Sanger sequencing is that the error rates for the called bases are markedly higher. This underlies the importance of obtaining high coverage 'depth' (the number of independent sequence reads aligned at one location). For this reason the removal of variants supported by only low read depth (e.g. 10 reads or less) is an important QC step.

Even given robust quality control throughout the analytical pipeline, the resulting file of SNPs and indels will contain many thousands of variants. The most pressing issue is how to determine the relationship (if any) of specific variants identified to the disease phenotype(s). Annotation of variants and filtering to identify and remove 'unimportant' variation can be achieved by tools such as Annovar (Wang et al. 2010) which enables local download of all variants in genomic databases (1,000 genomes, dbSNP, etc.) and provides tools for flexible filtering to remove common variation unlikely to be involved in disease. This is not straightforward since a number of these databases, such as recent versions of dbSNP, contain known rare and disease-causing variants which might be relevant to the phenotype under investigation. However, reduction in complexity of voluminous data at this stage is essential since an individual exome is likely to carry ~10,000 amino acid altering SNPs (Ng et al. 2010b). A (probably small) proportion of these are likely to negatively impact health, but the majority simply contribute to the large diversity of proteins and have little or no deleterious impact. For Mendelian diseases it is likely that the rare high-penetrance variants involved are private to affected individuals fully supporting the value of filtering out the common variation represented in genomic databases. Efficient filtering reduces the pool of potential disease influencing variants enabling cost-effective follow-up of a much smaller number of genes and/or variants. Studies of Mendelian disorders assume a single highly penetrant coding mutation is sufficient to cause disease and that mutation is very rare and probably restricted to affected individuals. The volume of variation can be much reduced by only considering variants that change the protein sequence (non-synonymous), coding indels and splice acceptor and donor site changes. However, for non-Mendelian traits, it is known, from GWAS studies, that common intronic, regulatory and synonymous variation has an impact on disease, and so filtering is likely to lose information. Even after filtration against common variant databases, and after considering only protein-changing variants, the high number of variants in an individual exome is large enough to challenge further progress. *In silico* approaches computationally evaluate potential disease severity of variants by making multispecies comparisons and using models of molecular evolution (Kumar et al. 2011). The degree of conservation at individual positions and databases of permitted substitutions indicates the potential impact of a given change. It is known that disease-associated SNPs are over-represented at locations in the genome that have changed to only a limited degree over evolutionary time. Variants at locations conserved throughout vertebrates are more likely to be involved in Mendelian disease, and the same has been found to be true for the locations of somatic variation in cancers. Intense purifying selection against damaging variants at these locations is likely to occur through a reduction in reproductive fitness. For this reason molecular evolutionary predictions are considered less useful for complex disease where later onset has limited impact on fecundity. However, there is a spectrum of genetic disease from single-gene Mendelian disorders to complex traits. Therefore, *in silico* prediction may be valuable for more 'extreme' forms of complex disease (e.g. early onset, more severe disease subtypes, familial cases). Ranking variants by their predicted or known effect on protein function and their degree of conservation using tools, such as SIFT (Kumar et al. 2009), PolyPhen2 (Adzhubei et al. 2010), LRT (Chun and Fay 2009) and MutationTaster (Schwartz et al. 2010), and composite databases of functional predictions such as dbNSFP (Liu et al. 2011) is an important further step towards reducing data depth and complexity. The various algorithms output scores which quantify the extent to which a non-synonymous variant is likely to be deleterious. Such an approach has already been used with success to prioritise novel variants for follow-up in Mendelian disease studies (Ng et al. 2010a). SIFT ('Sorting Tolerant From Intolerant', http://sift.bii.a-star.edu.sg/) predicts the effect on protein function of single amino acid changes. The SIFT algorithm works by searching for similar sequences that are likely to have matching functions, generates an alignment of those sequences and computes probabilities for all possible substitutions from the alignment. Those with $p<0.05$ are classified as deleterious mutations or, otherwise, tolerated. PhyloP (Pollard et al. 2010) similarly provides a conservation score highlighting locations that are conserved from invertebrates to humans in which substitutions are highly likely to disrupt critical protein function. PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/) also predicts the impact of an amino acid substitution on protein structure and function. The algorithm uses sequence and structural features to evaluate the impact of amino acid replacements within a multiple sequence alignment of homologous proteins, the extent of modification of the resultant protein, and whether the substituted allele originated at a particularly mutable site. The alignment process uses the set of homologous sequences and employs clustering to construct and refine their multiple alignment. The functional significance of a substitution is predicted from the set of features by a naive Bayes classifier (Adzhubei et al. 2010). Chun and Fay (2009) develop a likelihood ratio test (LRT, http://www.genetics.wustl.edu/jflab/lrt_query.html) which compares the null model of neutral codon evolution to the alternative model that the codon has evolved under negative selection. Deleterious mutations are considered to be the non-synonymous SNPs that significantly disrupt the constrained codons defined by the LRT. The LRT generates a p-value for the likelihood ratio test of codon constraint. The test is developed from data for 32 vertebrate species. Chun and Fay (2009) found, however, a disturbingly low degree of overlap between predictions made by the LRT, SIFT and PolyPhen with 76% of predictions unique to one of the three methods and only 5% of predictions made by all three. With this in mind Liu et al. (2011) argue that, because the various alternative algorithms have their own strengths and weaknesses, it is useful to construct a consensus prediction. This is presented in their dbNSFP database (http://sites.google.com/site/jpopgen/dbNSFP) which contains functional predictions from multiple algorithms compiling predicted scores for non-synonymous variants from SIFT, PolyPhen2, LRT, MutationTaster and PhyloP.

7.3.4 Collapsing Methods for Rare Variants in Large Samples

Rarer variants are likely to be enriched for alleles with functional disease impact and may show larger effect sizes than common alleles as a consequence of purifying selection. However, the penetrance of most of these variants is likely to be comparatively low (Bodmer and Bonilla 2008). Therefore, for most complex disease phenotypes, the cumulative impact of many rare variants is likely to contribute significantly to the disease phenotype. However, the power to detect such alleles is low due the relatively low penetrance, the small number of copies of a given variant present and the need for stringent correction for the number of variants tested. For this reason analytical approaches for large samples have been developed that test for the combined effects of a set of rare variants, thereby greatly reducing the number of statistical tests while maximising power. Such a 'collapsing' approach requires prior specification of the set of variants to be combined to make the test. Li and Leal (2008) point out that misclassification resulting from the collapsing of nonfunctional variants with functional sites adversely affects the power of the test. Misclassification can arise when noncausal variants are included and when functional variants are excluded because they either have not been sequenced or have incorrectly been classified as nonfunctional by bioinformatics tools. In contrast multiple-marker methods which test several sites for their influence on phenotype simultaneously are more robust to misclassification, but potentially less powerful than collapsing methods. Li and Leal's combined multivariate and collapsing (CMC) method aims to maximise power while being robust to misclassification. This and related tests are reviewed by Dering et al. (2011). The collapsing method defines an indicator variable *X* for the jth case individual to define whether or not that subject carries any rare variant in the target of interest (e.g. a gene) such that *X*j = 1 when a rare variant is present and 0 when absent with *Y*j similarly defined for controls. The test made is for association of multiple rare variants in which the proportion of rare variants in cases and controls differ. This is a fixed allele-frequency threshold approach for which power was investigated by Price et al. (2010). The authors examined different thresholds at which to define a variant as 'rare' (their T1 and T5 models representing 1 and 5% allele-frequency thresholds, respectively). They also describe a version of the test which weights (under the null hypothesis of no association) the contribution of each SNP by the inverse square root of the expected variance, based on allele frequencies computed from controls. This approach gives much higher weights to very rare variants. Price et al. propose a variable threshold approach which assumes an unknown threshold T for which variants with a MAF below the threshold are more likely to be functionally important than those above. The authors compute the maximum test statistic over a wide range of values of T to obtain the maximum of the threshold specific test statistics. The p-values are determined (as in all collapsing methods) by permutation tests.

An important addition to the range of collapsing approaches incorporates predicted functional information that improves the statistical test. Price et al. (2010) incorporated PolyPhen2 probabilistic scores for neutral and deleterious amino acid changes as weights in the regression. In their simulation study, setting the significance level to $p = 0.05$, power was higher at 60 and 69% for the variable threshold and variable threshold with PolyPhen scores models, respectively, compared to 55, 50 and 54% for the T1, T5 and weighted threshold models, respectively.

Luo et al. (2011) point out some of the limitations of collapsing methods, noting that variants at different genome locations may have different effect sizes which are unlikely to be determined only by their frequencies and collapsing without assigning weights that are functions of variant frequencies cannot fully exploit information of genetic effect sizes; multiple rare variants may be correlated, so grouping them needs to take this into account. They develop functional principal component analysis (FPCA)-based statistics for which they determine higher power to detect association with rare variants and enhanced ability to filter out sequence errors.

7.3.5 Copy Number Variant (CNV) and Loss of Heterozygosity Analysis

Test for structural variation has been typically undertaken using array comparative genome hybridization (CGH) which tests up to one million probes and can detect variants in the size range of 10–25 kilobases. But much higher resolution can be achieved from sequence data, and Yoon et al. (2009) develop methods for detecting CNVs in whole genome sequences. However, similar application to exome sequence data presents difficulty because the read sequence distribution is not random or unbiased and the read depths do not follow a normal distribution from which deviations suggest the presence of a copy number variant. However, if the biases are controlled, exome sequencing data present the opportunity to detect structural variants at much higher resolution and extend the utility of the data beyond the identification of single nucleotide variants and small indels. The problems presented by the discrete nature of the exome read distribution are considered by Sathirapongsasuti et al. (2011) who describe a method to detect copy number variations (CNVs) and loss of heterozygosity (LOH) in exome data. The approach uses normalised depth ratios in paired samples (such as tumour/germline) that have been processed in a similar way, including library preparation, and share similar average depth of coverage. This approach was shown to identify CNVs as small as 120 pb representing single exons with higher than average coverage. The read depth data can be more flexibly used in non-matched exome samples, for example, by using data from a pool of control exomes to serve as, effectively, a matched control sample (since the average copy number is likely to be two given a sufficiently large number of control exomes).

7.3.6 Strategies for Efficient Analysis and Data Management

The alignment of short sequence reads has been regarded as a major bottleneck in the analysis of NGS data (Li and Homer 2010). However, improving the algorithms and the development of tools which exploit distributed processors has reduced this bottleneck, at least for exome sequence. Important developments include platforms which automate pipelines and provide integration of bioinformatics tools to facilitate exome analysis. An example is Galaxy (Goecks et al. 2010, http://galaxy.psu.edu/) which provides a web-based platform to facilitate accessibility of NGS data analysis, exploiting the latest informatics tools, while tracking data provenance and ensuring reproducibility of analysis pathways undertaken. Galaxy is intended to free users from the necessity to develop computer code and the need to learn the implementation details of individual software packages. Galaxy offers a framework for performing exome studies which enables reconstruction of the analysis pathways undertaken by capturing details of analyses performed through a web interface. Perhaps most significant, given that that exome sequencing will shortly be superseded by far more challenging whole genome sequencing, is the development of a cloud computing enabled version (http://www.genomeweb.com/informatics/galaxy-joins-host-bioinformatics-projects-embracing-cloud-infrastructure-option). Cloud computing, in which computation is offered as a service, provides access to much greater computational power and storage than is available to an individual lab. Cloud computing is therefore regarded as a route to reducing some of the concerns about the management and analysis from the on-going and developing NGS ‘data deluge’.

7.4 Conclusions

A range of strategies are being employed to exploit exome sequencing for the identification of rarer variation underlying Mendelian disease and complex traits. Genotyping a small number of affected individuals in families showing strongly Mendelian patterns of inheritance has already proven to be a highly successful strategy with several important genes identified. Such an approach relies on the sharing of underlying causal variant(s) between family members. With higher penetrance variants, it is possible to combine evidence from linkage in these scenarios to reduce the list of potential causal targets. Thus, targeted follow-up can focus on the variants identified in these regions. For more complex phenotypes strategies include investigating cases with ‘extreme’ or otherwise unusual phenotypes (e.g. early onset disease, well-defined disease subtypes). Such an approach assumes that a relatively small number of moderate-penetrance variants might emerge as contributory to disease. In this situation family-based designs, where possible, are likely to reduce the overall complexity and number of targets for follow-up. Extensive filtration based on known or predicted gene function further delimits variants for greater consideration. From the study of cancer genomes, novel somatic variation can be identified by filtering out germline variation.

With respect to all studies involving complex disease in unrelated individuals, statistical analysis is plagued by low power and one strategy is to combine rare variants for analysis using some form of ‘collapsing’ approach.

In the longer term whole genome sequencing will replace exome sequencing and provides a range of new problems. The most obvious of these arises from the fact that it is now possible to produce DNA sequence more quickly and cheaply than the computing infrastructure can be developed to manage it (Stein 2010). Indeed the cost of sequencing is now decreasing much faster than the cost of storage of the data, and storage costs are likely to exceed the cost of production in the near future. Further development of novel strategies including cloud computing, in which hardware, runtime and data storage are effectively rented for specific projects, offers a credible way forwards. The Galaxy package has been implemented successfully on the Elastic Compute Cloud (EC2) web service offered by Amazon and provides a comprehensive range of cloud-enabled tools for NGS analysis. Such developments are promising although, as Stein (2010) points out, there remain major obstacles with respect to the network bandwidth and the transfer of huge volumes of data on and off networks. It is clear that the future development and application of NGS offers both great promise and major challenges.

References

Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

Alkan C, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7.

Bainbridge MN, et al. Whole exome capture in solution with 3Gbp of data. Genome Biol. 2010;11:R62.

Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695–701.

Burrows M, Wheeler D. A block sorting lossless data compression algorithm. Technical report 124. Palo Alto: Digital Equipment Corporation; 1994.

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.

Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11:415–25.

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40.

Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011;35:S12–17.

Feng B-J, et al. Design considerations for massively parallel sequencing studies of complex human disease. PLoS One. 2011;6(8):e23221.

Girard SL, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet. 2011;43(9):860–4.

Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:R86.

Harismendy O, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10:R32.

Howrigan DP, et al. Mutational load analysis of unrelated individuals. BMC Proc. 2011;5 Suppl 9:S55.

Johansen CT, et al. Mutation skew in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010;42(8):684–7.

Koboldt DC, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283–5.

Krawitz PM, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphoshatasia mental retardation syndrome. Nat Genet. 2010;42(10):827–9.

Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007;80(4): 727–39.

Kumar P, et al. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. Nat Protoc. 2009;4:1073–81.

Kumar S, Dudley JT, Filipski A, Liu L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends Genet. 2011;27(9):377–86.

Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187097.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

Lehne B, Lewis CM, Schlitt T. Exome localization of complex disease association signals. BMC Genomics. 2011;12:92.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754–60.

Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473–83.

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851–8.

Li H, et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009a;25:2078–9.

Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009b;25(15):1966–7.

Liu X, Jian X, Boerwinkle E. DbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894–9.

Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. Genome Res. 2011;21:1099–108.

Majewski J, Scwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011. doi:10.1136/jmedgenet-2011-100223.

McClellan J, King MC. Genetic heterogeneity and human disease. Cell. 2010;141:210–17.

Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6.

Ng SB, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010a;42:30–5.

Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. Hum Mol Genet. 2010b;19:R119–24.

Pireddu L, Leo S, Zanetti G. SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics. 2011;27(15):2159.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.

Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86:832–8.

Reis-Filho JS. Next-generation sequencing. Breast Cancer Res. 2009;11 Suppl 3:S12.

Rivas MA, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066–75.

Ruffalo M, LaFramboise T, Koyuturk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics. 2011. doi:10.1093/bioinformatics/btr477.

Rumble SM, et al. SHRiMP: accurate mapping of short color-space reads. PloS Comput Biol. 2009;5(5):e1000386.

Sathirapongsasuti JF, et al. Exome sequencing-based copy number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011. doi:10.1093/bioinformatics/btr462.

Schwartz JM, et al. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575–6.

Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11:207.

Tiacci E, et al. BRAF mutations in hairy-cell leukemia. N Engl J Med. 2011;364(24):2305–15.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Res. 2010;38:e164.

Yoon S, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19:1586–92.

Chapter 8
Rare Variants Analysis in Unrelated Individuals

Tao Feng and Xiaofeng Zhu

Abstract Although the genome-wide association studies, which are based on common disease-common variants (CDCV) hypothesis, have great success in dissecting the genetic architecture of human diseases, their limitation of explaining the missing heritability motivated researchers to test the hypothesis that rare variants contribute to the variation of common diseases, that is, common disease/rare variants (CDRV) hypothesis. The fast developed high-throughput next generation of sequencing technologies has made the studies of rare variants practicable. Statistical approaches to test associations between a phenotype and rare variants are rapidly developing. Overall, the key idea of these methods is to test a set of rare variants in a defined region or regions by collapsing or aggregating rare variants. To improve the statistical power, several weighting strategies to the rare variants and/or adding the informative covariates in the model have been published. In this chapter, some of these methods which can use unrelated individuals and family members are introduced.

Keywords GWAS • Common disease-common variants • Common disease rare variants • SNPs • Haplotype • Collapsing • Aggregation

8.1 Introduction

Genome-wide association studies (GWAS) have revealed significant evidence that specific common DNA sequence differences among people influence their genetic susceptibility to more than 60 different common diseases and created novel hypotheses for biological mechanism underlying complex diseases or traits.

T. Feng (✉) • X. Zhu
Case Western Reserve University, Cleveland, OH 44120, USA
e-mail: tao.feng@case.edu; xzhu1@darwin.EPBI.CWRU.edu

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_8, © Springer Science+Business Media Dordrecht 2012

However, it also raises some important questions on the roles of rare variants in human complex disease. The statistical methods commonly used in GWAS are typically underpowered to detect any effects of rare variants. In this review, we mainly focus on the rapidly developing methods to improve the statistical power for rare variants analysis; in particular, we described the methods with great details in the context of next generation sequencing data.

8.1.1 Success of GWAS and Its Limitation

With many investigators' effort in the last decades, our understanding of the genetic basis of disease risk has been improved greatly through genome-wide association studies (GWAS). The purpose of the GWAS is to uncover the connection between specific genes and their expression and then to expedite the identification of genetic risk factors for the development or progression of disease. To date, hundreds of GWAS have been performed to uncover the associate between particular genetic variations and diseases, such as hypertension, bipolar disease, coronary artery disease, diabetes, and cancer (Birney et al. 2007; Consortium WTCC 2007; Heid et al. 2010; Lango Allen et al. 2010). These eminent studies have successfully found thousands of genes which highly associate with hundreds of traits. As of 2nd quarter 2011, the US National Human Genome Resource Institute (NHGRI) GWAS catalogue lists 1,449 genome-wide significant associations with 237 traits and diseases spread across all auto chromosome except the Y chromosome (Hindorff et al.).

Originally, GWAS were designed as a genetic association study to capture a large proportion of the common variation in the human genome in a population by using the high-throughput genotyping technologies, and it was believed that the number of genotyped samples can provide sufficient power to detect variants of modest effect. However, GWAS, which is dominated by the simply statistical hypothesis common disease-common variants (CDCV), has challenged the missing heritability problem that the genetic variants identified by GWAS only account a small fraction of heritability observed in family studies (Manolio et al. 2009). For example, height is known to be a heritable trait with estimated heritability around 0.8 from family or twin studies, which implies about 80% of the individual variation and is attributable to genetic factors. Although the three GWAS in 2008 (Gudbjartsson et al. 2008; Lettre et al. 2008; Weedon et al. 2008) identified 40 previously unknown variants, each one only explains 0.3–0.5% of the phenotypic variance. The results from GWAS suggest that there must be genetic factors contributing to common complex diseases that are simply not amenable to detection via the GWAS strategy (Pritchard 2001). Some researchers argue that the missing heritability may be accounted for by many rare variants with relatively large effective sizes, or interactions, such as gene-gene or gene-environment interactions (Bansal et al. 2010; Manolio et al. 2009; Zuk et al. 2012).

8.1.2 Detecting Rare Variants

There has been growing debate over the nature of the genetic contribution to individual susceptibility to common complex. Comparing with the CDCV, common disease rare variant (CDRV) argues that multiple rare DNA sequence variants are major contribution of genetic susceptibility to common disease. Differing from that, a common variant usually has modest or low disease penetrance; a rare variant has relatively large disease penetrance. With the new sequencing technologies and publication of 1000 Genomes Project (2010), we are at the era that can test the CDRV hypothesis. By directly testing many rare variants on candidate genes, these studies have identified collections of rare variants associated with phenotypic variation, such as multiple functional variants in *IFIH1*, *NPC1L1*, *PCSK9*, *SLC12A3*, *SLC12A1,* and *KCNJ1* associated with type I diabetes, sterol absorption, plasma levels of LDL-C, and blood pressure (Cohen et al. 2005, 2006; Ji et al. 2008; Nejentsev et al. 2009).

In the following section, we will introduce statistical methods for testing rare variant association that can be applied for unrelated individuals.

8.2 Data Description and Methods

Below we first (in Sect. 8.2.1) describe the data structure of any genetic variants either located in a candidate gene or a genomic region and clearly define the relevant parameters. We then exhaustively review all of the previously published methods focusing on statistical collapsing (between Sects. 8.2.2.1 and 8.2.2.11).

8.2.1 Data Describe

Assume we test the association of genetic variants and disease status in a candidate gene or region, which includes L SNPs in it, and total N unrelated individuals with either quantitative traits or binary traits being collected. Further, let y_i denote the quantitative trait or binary trait and use A_j and a_j denote the two alleles of jth SNP, in which A_i always refers to the rare allele and has an allele frequency p_i. Furthermore, let code $x_{ij} = 0,1$ or 2 be the number of minor alleles at the jth SNP carried by the ith individual, where $i = 1,\ldots,N$ and $j = 1,\ldots,L$.

8.2.2 Methods

Between Sects. 8.2.2.1 and 8.2.2.11, we will provide mathematical details for each method illustrated and will also provide our insights of specific merits and limitations of each method.

8.2.2.1 Collapsing Method

In contrast to common variants, the power of traditional statistical methods to detect rare variant association is usually poor and requires large sample sizes due to the small minor allele frequencies (MAF) of rare variants. Although a rare variant individually may make only a tiny contribution to a phenotypic variation, collectively rare variants may uncover a substantial proportion of missing heritability (Gibson 2010; Manolio et al. 2009). Based on this principle, collapsing method has been proposed to improve statistical power for a binary trait. To do this, we define an indicator variable G_i for the *i*th individual as $G_i = 1$ if rare variant(s) is(are) present, otherwise as $G_i = 0$. The detection of an association of multiple rare variants is transformed into a test of whether the proportions of individuals with rare variants in cases and controls differ. Then any single SNP association test that is applied in GWAS can be applied here, such as a chi-squared test for a contingence table or a regression analysis. In 2006, Cohen et al. suggested a method to compare the number of rare variants unique to either cases or controls using Fisher's exact test (Cohen et al. 2006). This method is simple and fast, but it has its limitation. If the number of SNPs in a considered region is large, it has more chance that variable G_i will be coded as 1, and then there will be little difference between cases and control, resulting poor statistical power. One way to improve this method is considering the number of rare variants presented in an individual rather than simply coding 1 or 0 for the individual. Another way is partitioning the region into several small regions and then use multivariate test proposed by Li and Leal (2008).

8.2.2.2 Combined Multivariate and Collapsing

To take advantage of both the multiple marker tests and the collapsing method, Li and Leal (2008) considered an extension of the collapsing method, which they termed the combined multivariate and collapsing (CMC) method. For a considered region, they first divide the markers (e.g., SNPs) within the region into groups according certain criteria (e.g., allele frequencies) and then collapse the rare variants within each group using the method describe in Sect. 8.2.2.1. To analyze the groups of collapsed rare variants, a multivariate test such as Hotelling's T^2 test is applied.

The shortcoming of this method is that the power will decrease when the number of subgroups increases. The criteria of the partition also can affect the power of the test. Furthermore, collapsing method assumes that each rare variant has the same contribution to the disease susceptibility and this may not be true in reality.

8.2.2.3 Weighted Sum Association Method (WSM)

Madsen and Browning (2009) proposed a statistic for testing a prespecified collapsed set of variants that weights each variant by its frequency, thus allowing one to include variants of any frequencies into the collapsed set. This approach proceeds by defining the genetic score of individual i as $\gamma_i = \sum_{j=1}^{L} \frac{x_{ij}}{w_j}$, where a nonzero w_j is the weight of jth variant and is defined by $w_j = \sqrt{Np_j(1-p_j)}$. Madsen and Browning suggested that the MAF p_j is estimated by controls only. Thus, for individual i, γ_i represents a single core that is obtained by combining information from all the L variants in the region of interest. An association test is performed by testing this score rather than testing the individual variants. Madsen and Browning (2009) suggest using a nonparametric Wilcoxon's test for the association test and calculating the p-value using a permutation approach.

When the interesting region includes multiple common variants, Feng et al. (2011) suggest the power of WSM will decrease. Although Madsen and Browning (2009) did not suggest using a threshold model, a predefined threshold α, such that the weight will be 0 if a variant with $MAF > \alpha$ and this SNP will be exclude from the test, can often improve the power when only rare variants are associated with a disease status. However, it is difficult to select an optimal threshold in practice. Price et al. (2010) proposed a variable-threshold approach for testing rare coding variants to solve this problem.

8.2.2.4 Pooled Association Tests for Rare Variants

To obtain the optimal MAF threshold α for which variants with a MAF below α are substantially more likely to be functional than are variants with an MAF above α, a data-driven z-score $z(\alpha)$ for each allele-frequency threshold α is computed, and the maximum z-score across different values of α is defined as zMax. Then a permutation procedure is used to assess the statistical significance of zMax, allowing zMax in the permuted data to be attained at values of α different from those in un-permuted data to ensure the validity of the permutation test. We refer the reader to Price et al. (2010) for details about the calculation of the z-scores and for testing the statistical significance of the variants using this method.

Besides finding the optimal cutoff of MAF α, Price et al. also proposed using the functional relevance of the individual variants to define the weights. They suggest using the PolyPhen-2 scores (Ramensky et al. 2002; Adzhubei et al. 2010), which evaluate the possible functional effect of an SNP by calculating the distributions of PolyPhen-2 probabilistic scores for neutral and damaging amino acid changes. We refer the reader to Price et al. (2010) for details about this method.

8.2.2.5 A Data-Adaptive aSum Test (Consider the Direction)

Han and Pan (2010) proposed a data-adaptive modification to sum test and aimed to strike a balance between utilizing information on multiple markers in linkage disequilibrium and reducing the cost of large degrees of freedom or of multiple testing adjustment. For the rare variants, the logistic regression model

$$\text{Logit}\,\Pr(y_i = 1) = \beta_{c0} + \sum_{j=1}^{L} x_{ij}\beta_c$$

is applied to test any possible association between the disease and SNPs. Under null hypothesis $H_0 : \beta_c = 0$, the test statistics has an asymptotic χ^2 distribution with 1 degree of freedom (DF). The main advantage of this sum test is that, because it tests on only one parameter β_c, there will be no power loss due to large DF or multiple testing adjustments. However, the test may have reduced power with a small $\hat{\beta}_c$, the maximum likelihood estimate of β_c, when the SNPs have different directions of contribution, that is, some of variants in the region are harmful and others are beneficial. The data-adaptive sum test (aSum) adapts the coding x_{ij} of each SNP *j* by adding a sign based on the estimated coefficient of logistic regression of SNP *j*.

Furthermore, they modify aSum test to combine the rare variants into one group and the common variants into another group by summing over their genotypic coding, then test on the two corresponding regression coefficients in a logistic regression model (termed aSumC test). There are two potential advantages of this method. First, this test can overcome the problem with different association directions of the functional variants, from which both the CMC and the WSM tests suffer with possibly significant power loss. Second, with only two groups, the aSumC may have a much smaller number of DF and thus higher power than the CMC test.

Hoffmann and Witte (Hoffmann et al. 2010) proposed a general framework of the aSum test by adding the weight w_i in the logistic regression, that is, $g(y_i) = \alpha_0 + \gamma\left[\sum_{j=1}^{L} w_j x_{ij}\right]$, where g is the link function and the weight w_j is define by $1/\sqrt{p_j(1-p_j)}$, similar as the Madsen and Browning's weight.

8.2.2.6 Alpha Test

C-alpha is a well-established and powerful test for the presence of a mixture of biased and neutral coins (Neyman and Scott 1966; Zelterman and Chen 1988). Neale et al. (2011) tailored the C-alpha score test and applied it to test a set of rare variants for association. Under the assumption that the rare variants are distributed at random across the subjects, the binomial (n, p) distribution evaluates the probability of observing a particular variant y times in the cases out of n total. Under the balanced sample of cases and controls, in other word that $p = 0.5$, the y to be 0,1 and 2 for $n = 1$ are expected with probability 0.25, 0.5, and 0.25, respectively. If some variants are causal, the higher proportion of doubletons with $y = 2$ and/or $y = 0$ is expected. Due to each variant cannot provide sufficient information to draw a firm conclusion about the association, the C-alpha test was applied to detect a pattern across the full set of rare variants in the target region.

In detail, for the *j*th variant, assume y_j is a binomial (n_j, p_j) if the rare variants was observed n_j times. Under the null hypothesis, $p_j = p_0$ (say 0.50 if cases and controls are equal in number), and under the alternative hypothesis, p_j follows a mixture distribution across the L variants with some variants detrimental ($p_j > p_0$), some neutral, and some protective ($p_j < p_0$). The C-alpha test statistic

$$T = \sum_{j=1}^{L} [(y_j - n_j p_0)^2 - n_j p_0 (1 - p_0)],$$

contrasts the variance of each observed count with the expected variance. The variance of T is derived by

$$c = \sum_{n=2}^{\max n} m(n) \sum_{u=0}^{n} [(u - np_0)^2 - np_0(1 - p_0)]^2 f(u \mid n, p_0),$$

where $m(n)$ is the number of variants with n copies and $f(u \mid n, p_0)$ denotes the probability of observing u copies of the *i*th variant assuming the binomial model. The resulting test statistic is defined as $Z = \frac{T}{\sqrt{c}} \sim N(0,1)$. The null hypothesis will be rejected when Z is larger than expected based on a one-tailed standard normal distribution.

The C-alpha test is a non-burden-based test and is hence robust to the direction and magnitude of effect, and this allow the C-alpha test to have improved power over other burden-based tests, especially when the effects are in different directions. But the covariate is not easier to be adjusted in the C-alpha. Also, the C-alpha test uses permutation to obtain a p-value when linkage disequilibrium is present among the variants, and the approach also has not been generalized to analysis of quantitative trait.

8.2.2.7 Sequence Kernel Association Test (SKAT)

Wu and Lin (Wu et al. 2011) introduced the kernel function into the regression model and combine the SNPs in the considered region with linear or nonlinear weights. The sequence kernel association test (SKAT) extends kernel machine-based tests for rare variants with more accurate asymptotic approximations in the tail distribution. This method is supervised for the joint effects of multiple variants in a region on a phenotype; it is flexible and computationally efficient to test for association between genetic variant in a region and a continuous or dichotomous trait while easily adjusting for covariates.

The SKAT test starts with a linear model

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z}_\mathrm{i} + \boldsymbol{\beta}'\mathbf{X}_\mathrm{i} + \varepsilon_i,$$

when the phenotype are continuous traits, and the logistic model

$$\text{logit}\, P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z_i} + \boldsymbol{\beta}'\mathbf{X_i},$$

when the phenotype are binary traits (e.g., $y=0/1$ for case or control). Here, $\mathbf{Z_i} = (z_{i1}, z_{i2}, \ldots, z_{im})$ denotes the covariates, $\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{iL})$ denotes the genotypes for the L variants within the region, α_0 is an intercept term, $\boldsymbol{\alpha}' = [\alpha_1, \ldots, \alpha_m]'$ is the vector of regression coefficients for m covariates. $\boldsymbol{\beta}' = [\beta_1, \ldots, \beta_L]'$ is the vector of regression coefficients for the L observed gene variants in the region, and for continuous phenotypes, ε_i is an error term with a mean of zero and a variance of σ^2.

Under the null hypothesis $H_0 : \beta = 0$ or $\beta_1 = \beta_2 = \cdots = \beta_L = 0$, the standard L-DF likelihood ratio test has little power. Given the additional assumption that each β_i follows an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance component and w_j is a prespecified weight for variant j, the SKAT can improve the power by testing $H_0 : \tau = 0$. To do the test, the variance-component score statistic

$$Q = (\mathbf{y} - \hat{u})' \mathbf{K} (\mathbf{y} - \hat{u})$$

is applied. In the above formula, $\mathbf{K} = \text{XWX}'$, $\hat{u}$ is the predicted mean of $\mathbf{y}$ under H_0, that is, $\hat{u} = \hat{\alpha}_0 + \mathbf{Z}\widehat{\alpha}$ for continuous traits and $\hat{u} = \text{logit}^{-1}(\hat{\alpha}_0 + \mathbf{Z}\widehat{\alpha})$ for dichotomous traits, and $\hat{\alpha}_0$ and $\widehat{\alpha}$ are estimated under the null hypothesis by regressing $\mathbf{y}$ on the covariates $\mathbf{X}$ only. Here, $\mathbf{X}$ is an $N \times L$ matrix with the $(i, j)-$th element being the genotype of *j*th variant in *i*th individual, and $\mathbf{W} = diag(w_1, \ldots, w_L)$ contains the weights of the L variants. $\mathbf{K}$ is an $N \times N$ matrix with the $(i, i')-$th element equal to $K(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{j=1}^{L} w_j X_{ij} X_{i'j}$. $K(\bullet, \bullet)$ is called the weighted linear kernel function, and $K(\mathbf{X}_i, \mathbf{X}_{i'})$ measures the genetic similarity between individual i and i' in the region via the L markers. An attractive feature of SKAT is the ability to model the epistatic effects of sequence variants on the phenotype within the flexible kernel machine regression framework. To do so, the term $\boldsymbol{\beta}'\mathbf{X_i}$ was replaced by a more flexible function $f(\mathbf{X_i})$ in the linear and logistic model. $f(\mathbf{X_i})$ allows for the interactions of rare variant by rare variant or common variant by rare variant. For the purpose of rare variant analysis, the weighted quadratic kernel can be chosen as $K(\mathbf{X}_i, \mathbf{X}_{i'}) = (1 + \sum_{j=1}^{L} w_j X_{ij} X_{i'j})^2$ or the weighted identity by state (IBS) kernel $K(\mathbf{X}_i, \mathbf{X}_{i'}) = \sum_{j=1}^{L} w_j IBS(X_{ij}, X_{i'j})$. A question is how to choose w_j in the kernel function, which can affect statistical power. Wu et al. (2011) suggested $\sqrt{w_j} \sim Beta(MAF_j; a_1, a_2)$, the beta distribution function with prespecified parameters a_1 and a_2 evaluated at the sample MAF using both cases and controls for the *j*th variant in the data. The setting $a_1 = 1$ and $a_2 = 25$ was suggested because it increases the weight of rare variants while still putting decent nonzero weights for variants with MAF 1–5%. When the outcome is dichotomous, no covariates are included and all $w_i = 1$; the SKAT test statistic Q is equivalent to the C-alpha test statistic T. Hence, SKAT can be seen as a generalized C-alpha test that does not require permutation but calculates the p-value analytically, allows for covariate adjustment, and accommodates either dichotomous or continuous phenotypes.

8.2.2.8 A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies

Lin and Tang (2011) also proposed a so-called general framework for association testing with rare variants by combining mutation information across multiple variant sites within a gene and relating the enriched genetic information to disease phenotypes through appropriate regression models. This framework in theory covers all major study designs (i.e., case-control, cross-sectional, cohort and family studies) and all common phenotypes (e.g., binary, quantitative, and age at onset), and it allows the incorporation of arbitrary covariates (e.g., environmental factors and ancestry variables).

Using the predefined notation, the logistic regression model logit $P(y_{\mathrm{i}}=1) = \boldsymbol{\alpha}'\mathbf{Z_i} + \boldsymbol{\beta}'\mathbf{X_i}$ is applied here, where vector $\mathbf{Z_i} = (1, z_{i1}, z_{i2}, \ldots, z_{im})$ denotes the m covariates. Let $\beta = \tau\xi$, where τ is a scalar constant and $\xi = \beta / \tau$. Then the logistic regression model becomes

$$\operatorname{logit}\Pr(y_i = 1) = \tau S_i + \boldsymbol{\gamma}'\mathbf{Z_i},$$

where $S_i = \boldsymbol{\xi}'\mathbf{X_i}$. Note that $\xi = (\xi_1, \ldots, \xi_L)'$ is a $L \times 1$ vector of weights and that S_i is a weighted linear combination of $x_{i1}, x_{i2}, \ldots x_{iL}$ with x_{ij} receiving the weight ξ_j. Here, ξ is referred as the weight function. The score statistic for testing the null hypothesis $H_0 : \tau = 0$ takes the form

$$U = \sum_{i=1}^{N} \left(y_i - \frac{e^{\hat{\gamma}' Z_i}}{1 + e^{\hat{\gamma}' Z_i}} \right) S_i,$$

where $\hat{\gamma}$ is the restricted maximum likelihood estimator of γ and solves the equation $\sum_{i=1}^{N} \left(y_i - \frac{e^{\hat{\gamma}' Z_i}}{1 + e^{\hat{\gamma}' Z_i}} \right) Z_i = 0$. The variance of U is estimated by

$$V = \sum_{i=1}^{N} v_i S_i^2 - \left(\sum_{i=1}^{N} v_i S_i Z_i \right)' \left(\sum_{i=1}^{N} v_i Z_i Z' \right)^{-1} \left(\sum_{i=1}^{N} v_i S_i Z_i \right),$$

where $v_i = \frac{e^{\hat{\gamma}' Z_i}}{(1 + e^{\hat{\gamma}' Z_i})^2}$. Under H_0, the test statistic $T = U / V^{1/2}$ is asymptotically standard normal. In the absence of covariates,

$$U = \sum_{i=1}^{N} (y_i - \bar{y}) S_i, \text{ and } V = \bar{y}(1 - \bar{y}) - \left\{ \sum_{i=1}^{N} S_i^2 - \frac{1}{N} \left(\sum_{i=1}^{N} S_i \right)^2 \right\},$$

where $\overline{y} = N^{-1}\sum_{i=1}^{N} y_i$.

Since the setting of weight function $\xi = (\xi_1,\ldots,\xi_L)'$ is unknown and must be determined biologically or empirically, several considerations were discussed:

1. If the choice of weight function ξ or the limit of the estimate of ξ is proportional to β, then the statistic T is the most powerful among all valid tests. Otherwise, U is no longer the score statistics. But it can be proved that statistic T is asymptotically standard normal under H_0 regardless how ξ is chosen.
2. This method allows not only for multiple allele-frequency thresholds but also for different types of weight functions. It can be shown that for K choices of ξ, which could correspond to different thresholds or different types of weight functions or both, the maximum of the absolute test statistics $T_{\max} = \max_{k=1,\ldots,K} |T_k|$ is applied, where the test statistics $T_k = U_k / V_k^{1/2}$ is defined for the kth choice of ξ. The score statistics U_k and its variance in the test statistics T_k are defined by $U_k = \sum_{i=1}^{N}\left(Y_i - \frac{e^{\hat{\gamma}'Z_i}}{1+e^{\hat{\gamma}'Z_i}}\right)S_{ki}$ and $V_k = \sum_{i=1}^{N} v_i S_{ki}^2 - \left(\sum_{i=1}^{N} v_i S_{ki} Z_i\right)\left(\sum_{i=1}^{N} v_i Z_i Z_i^1\right)^{-1}\left(\sum_{i=1}^{N} v_i S_{ki} Z_i\right)$ with corresponding kth S_i denoted by S_{ki}. If $t_{\max}$ would be the observed value of $T_{\max}$, then the p-value is given by

$$\Pr(T_{\max} > t_{\max}) = 1 - \Pr(|T_1| < t_{\max},\ldots,|T_k| < t_{\max}),$$

which is evaluated by treating $(T_1,\ldots,T_K)'$ as a K-variant normal random vector with mean 0 and a covariance matrix of $\{r_{kl}; k,l = 1,\ldots,K\}$, where $r_{kl} = V_{kl} / (V_{kk}V_{ll})^{1/2}$, $V_{kl} = \sum_{i=1}^{N} U_{ki}U_{li}$, and $U_{kl} = \left(y_i - \frac{e^{\hat{\gamma}Z_i}}{1+e^{\hat{\gamma}Z_i}}\right)\left\{S_{ki} - \left(\sum_{i=1}^{N} v_i S_{ki} Z_i\right)\left(\sum_{i=1}^{N} v_i Z_i Z_i'\right)^{-1} Z_i\right\}$. The H_0 will be rejected if the p-value is smaller than the nominal significance level α.
3. If set $\xi_j = 1 (j = 1,\ldots,L)$, then statistic T is a burden test. If it is sure that common variants are not associated with the phenotype, then setting $\xi_j = 0$ if MAF of jth SNP $p_j > c$, where c is a prespecified threshold (such as c=0.02 or 0.01). If setting $\xi_j = \{p_j(1-p_j)\}^{-1/2} (j = 1,\ldots,L)$, then the weight function is the same as that of Madsen and Browning. Differing from the Madsen and Browning's and Price et al. method, this method does not need permutation when sample is large enough. This method can also accommodate covariates and the result holds for all phenotypes. In addition, the SKAT statistic can be written as $Q = \sum_{j=1}^{L} \xi_j U_j^2$, where U_j is the jth component of the score statistic for testing the null hypothesis $\beta = 0$ in the above defined logistic regression model. The C-alpha statistic of Neale et al. (2011) is a special case of Q with $\xi_j = 1$ for binary traits without

covariates. If statistic U is rewritten as $\sum_{j=1}^{L} \xi_j U_j$, Han and Pan (2010) statistic is a special case of U (for binary traits without covariates) in which $\xi_j = -1$ if $\hat{\xi}_j < 0$, and the corresponding p-value <0.1 and $\xi_j = 1$ otherwise.

8.2.2.9 Haplotype-Based Collapsing Test

Besides directly using the genotype to collapse the rare variants, comparing haplotype frequencies between cases and controls (Zhu et al. 2010; Guo and Lin 2009; Li et al. 2010; Zhu et al. 2005, 2010) is another way to analyze the rare variants. These methods assume that the haplotypes created by the common and rare variants are able to tag multiple rare ungenotyped variants. Since very rare variants are usually not well tagged by common variants (Durbin et al. 2010), the haplotype-based methods may only work for identifying rare variants with MAF>0.5% (Li et al. 2010).

We introduce the two-stage approach here. At the first stage, a set of susceptibility haplotypes is identified by comparing their frequencies between cases and controls using a subset of samples. At the second stage, the cumulative susceptibility haplotype frequencies are compared using the rest of samples.

In detail, assuming total N individuals of whom n^u are unaffected (controls) and the remaining $N - n^u$ are affected (cases). At stage 1, we randomly select n $(< n^u)$ unaffected and m $(< N - n^u)$ affected individuals. We assume that the disease is rare and that the unaffected individuals are representative of the general population. Assume there are k different haplotypes $h_1, h_2, \ldots, h_k$ with observed haplotype frequencies $p_1, p_2, \ldots, p_k$ in the selected cases. Correspondingly, the ith haplotype has haplotype frequency p_i^0 in the controls. Then the risk haplotype set is defined as

$$S = \left\{ h_i \,\middle|\, p_i - p_i^0 > \gamma \sqrt{\frac{p_i^0 \left(1 - p_i^0\right)}{2n}} \right\},$$

where $\lambda = 1.28$ or 1.64 is a predefined number that affects the misclassification rate and power.

It has been demonstrated that rare risk haplotypes can be enriched in affected sibpairs (Zhu et al. 2010), and this information can be used to define risk haplotype as using unrelated individuals. When we have affected sibpairs available, we can define risk haplotype set using affected sibpairs. Assume there are M affected sibpairs and the haplotypes have been inferred, then the rare risk haplotype set for affected sibpairs can be defined by

$$S = \left\{ h_i \,\middle|\, p_i - p_i^0 > \gamma \sqrt{\frac{p_i^0 \left(1 - p_i^0\right)}{3M}} \right\}$$

where h_i, p_i, p_i^0 are the haplotype, its frequency in affected sibpairs and controls, respectively, and γ is defined as before. Here we used $3M$ because there are only $3M$ independent haplotypes in M sibpairs under the null hypothesis.

At the second stage, we test association of the risk set of haplotypes defined at stage 1 using the remaining $n^u - n$ unaffected individuals and the $N - n^u - n$ affected individuals. We compare the sum of the risk haplotypes frequencies in the cases and controls by Fisher's exact test. The weighted sum test, which is an extension of the two-stage method, was studied by Li et al. (2010).

To apply haplotype-based methods, haplotype phases have to be inferred, which adds a substantial computational burden. However, since we only need to infer the haplotype phases once in any data analysis, the computation is still within feasible limits. When risk variants are extremely rare (<0.5%), the power of haplotype-based methods can be low.

8.2.2.10 Odds Ratio Weighted Sum Statistic (ORWSS)

Price et al. (2010) demonstrated that the weights by Madsen and Browning (2009) are proportional to the log odds ratio for a variant. In addition, a coefficient in a logistic regression is equivalent to the logarithm of the corresponding odds ratio. Feng and Zhu (Feng et al. 2011) proposed a method, for the binary trait, which directly uses the odd ratio of a variant as the weight for that variant, rather than the variance estimated in controls. That is, the odds ratio between allele *A* at the *j*th SNP and a disease status using a 2×2 table was calculated. Since only rare variants are interested in and the corresponding 2×2 table may consist of entries with 0 observations, the amended estimator of the odds ratio by adding 0.5 to each cell was applied. It has been suggested that the amended estimator of the odds ratio behaves well (Agresti 2002). Then, let γ_j denote the logarithm of the amended odds ratio testing for the association of allele *A* at the *j*th SNP using all the cases and controls.

If y_i is a quantitative trait, the estimated coefficient γ_j of a linear regress model $y_i = \gamma_0 + x_{ij}\gamma_j + \varepsilon_i$ can be used as the weight for the *j*th variant. In detail, the weight of the *j*th SNP is defined as $\hat{\gamma}_j = \left(X_j' X_j\right)^{-1} X_j' Y = \dfrac{\sum_{i=1}^{N}\left(x_{ij}y_i - \bar{x}_j\bar{y}\right)}{\sum_{i=1}^{N}\left(x_{ij}^2 - \bar{x}_j^2\right)}$, where $\bar{x}_j$ and $\bar{y}$ are the mean of *j*th SNP and quantitative trait *Y*, respectively.

For the rare variants, the estimated coefficient $\hat{\gamma}_j$ may vary widely if the sample size is not large enough. Based on this consideration, the weight can be defined by $\dfrac{\hat{\gamma}_{\mathrm{j}}}{sd}$, where sd is the standard error of $\hat{\gamma}_j$.

The power of the existing rare variant methods is dependent on the threshold used to define a rare variant, which can result in misspecification of risk variants by either including neutral variants or excluding risk variants (Zawistowski et al. 2010). Price et al. (2010) addressed this issue via a variable MAF threshold at the cost of more computation. This problem can be worse for these pooling methods when both common and rare variants contribute to disease risk. When the MAF threshold is increased, many common neutral variants are also included – resulting in a dilution of association evidence. To overcome this limitation, the weight for the *j*th SNP is defined as

$$w_j = \begin{cases} \gamma_j, & \text{if } \gamma_j > \overline{\gamma} + c\sigma \text{ or } \gamma_j < \overline{\gamma} - c\sigma \\ 0, & otherwise \end{cases},$$

where σ is the standard deviation calculated from $\gamma_j, j = 1, \ldots L,$ $c = 1.64$ or 1.28 is a parameter, and $\overline{\gamma}$ is the mean. After defining the weight in this way, a same test procedure as Madsen and Browning's can be applied for the association test.

8.2.2.11 Combining Related and Unrelated Individual Together to Detect Rare Variants

Previously, it was demonstrated that rare risk variants will be enriched in ascertained families such as affected sibpairs (Zhu et al. 2010). Here, we illustrate how to use families, such as affected sibpairs or discordant sibpairs, to define the weights. Then a same test procedure as Madsen and Browning's test can be applied to do the association test. This method was called sibpair-based weighted sum statistic test (SPWSS) and it has been shown that with the same size of genotype effect, using family data can greatly increase statistical power in detecting rare risk variants (Feng et al. 2011). Here, the assumption that a minor allele is either a risk allele or neutral was made, but the similar methods can be applied to detect protective variants.

(1) Affected Sibpair Design

Assume there are N_{sib} affected sibpairs and *L* SNPs in the region. Further assume a SNP has two alleles *A* and *a*, and *A* always refers to the minor allele for all the SNPs as defined before and let ~ represent either the *A* or *a* allele at any SNP. Denote the *i*th sibpair's genotypes of the *L* SNPs as $g_i = \left((g_{i11}, g_{i21}), (g_{i12}, g_{i22}), \ldots, (g_{i1L}, g_{i2L})\right)$ where (g_{i1j}, g_{i2j}) refers to the *j*th SNP's genotypes for the *i*th sibpair. There is no need to differentiate the first or second sib here. The idea here is that if *A* at the *j*th SNP is a risk allele, the weight for this allele *A* should be proportional to the ratio of the risk from both affected sibpairs carrying A to that in general population. If this is the case, the weight will only depend on the alleles carried at the *j*th SNP. To do this, two scenarios are considered. First, if both affected sibs carry *A* at the *j*th SNP, the weight of *A* at this SNP is proportional to

$$\frac{Pr\left(both\ sibs\ are\ affected \mid (g_{i1j}, g_{i2j}) = (A\sim, A\sim)\right)}{p(both\ sibs\ are\ affected)}$$
$$= \frac{Pr\left((g_{i1j}, g_{i2j}) = (A\sim, A\sim) \mid both\ sibs\ are\ affected\right)}{\phi_I}$$

where $\phi_1 = Pr\left((g_{i1j}, g_{i2j}) = (A\sim, A\sim)\right)$. Second, if one sib carries A at the jth SNP and the other does not, the weight of A is dependent on how many other sites have an A allele carried by the other affected sib. That is, the weight is proportional to

$$\frac{Pr\left[\mathit{both\ sibs\ are\ affected} \mid (g_{i1j}, g_{i2j}) = (A\sim, aa), A\ \mathit{present\ at\ other\ sites} \text{ of sib} 2\right]}{p(\mathit{both\ sibs\ are\ affected})}$$
$$= \frac{Pr\left[(g_{i1j}, g_{i2j}) = (A\sim, aa), A\ \mathit{present\ at\ other\ sites} \text{ of sib} 2 \mid \mathit{both\ sibs\ are\ affected}\right]}{\phi_2}$$

where $\phi_2 = P\left[(g_{i1j}, g_{i2j}) = (A\sim, aa), A\ \mathit{present\ at\ other\ sites} \text{ of sib} 2\right]$. In the above equation, we always assume the first sib carries allele A when one of the two sibs carries allele A at the jth marker for easy description.

Based on above equations, a genotype score for each SNP in a sibpair can be defined. To do so, the genotype score of the jth SNP carried by ith affected sibpair was defined as

$$\tilde{g}_{ij} = \begin{cases} \dfrac{1}{\phi_2}, \text{ when } (g_{i1j}, g_{i2j}) = (A\sim, A\sim) \\ \dfrac{L_0}{2L\phi_2}, \quad \text{when } (g_{i1j}, g_{i2j}) = (A\sim, aa), \text{and } L_0 \mathit{of\ the\ other\ SNPs\ carry\ an\ A} \\ \qquad \mathit{allele} \text{ for sib 2} \\ 0, \quad \text{otherwise} \end{cases}.$$

In the above equation, the second term was divided by 2 because either one of the sibs may carry the $A\sim$ genotype at the jth SNP. The formulas for calculating ϕ_1 and ϕ_2 are given by

$$\begin{aligned} \phi_1 &= Pr\left((g_{i1j}, g_{i2j}) = (A\sim, A\sim)\right) \\ &= \sum_{I=1}^{2} Pr\left[(g_{i1j}, g_{i2j}) = (A\sim, A\sim) \mid I\right] \\ &= p_j\left(1 + \tfrac{1}{4}p_j - \tfrac{1}{4}p_j^2\right) + \tfrac{1}{4}p_j^2\left[1 + 2p_j(1-p_j) + 3(1-p_j)^2\right] \end{aligned}$$

and

$$\begin{aligned} \phi_2 &= Pr\left[(g_{i1j}, g_{i2j}) = (A\sim, aa), A\ \mathit{present\ at\ other\ sites} \text{ of sib} 2\right] \\ &= Pr\left[(g_{i1j}, g_{i2j}) = (A\sim, aa)\right]\left(1 - P[A \text{ not } \mathit{present\ at} \text{ any } \mathit{sites}]\right) \\ &= \sum_{I=1}^{2} Pr\left[(g_{i1j}, g_{i2j}) = (A\sim, aa) \mid I\right] P(I)\left(1 - P[A \text{ not } \mathit{present\ at} \text{ any } \mathit{sites}]\right. \\ &= \tfrac{1}{4}p_j\left(1-p_j\right)^2\left(4-p_j\right)\left[1 - \prod_{k \neq j}^{L}\left(1-p_k\right)^2\right] \end{aligned}$$

where p_j is the A allele frequency at the *j*th SNP which is estimated only in controls. There is also an assumption that all SNPs are in linkage equilibrium for obtaining ϕ_2. This may not be a reasonable assumption in the real data. However, simulations suggest that this assumption has little effect on the testing results (Feng et al. 2011).

For the *j*th SNP, we then calculate $\gamma_j = \frac{1}{N_{sib}} \sum_{i=1}^{N_{sib}} \tilde{g}_{ij}$, which is the average of the genotype scores across whole affected sibpairs. Under the alternative hypothesis, in which only a subset of variants are risk variants, we would expect these variants to be outliers. We thus define the weight for the *j*th SNP to be

$$w_j = \begin{cases} \gamma_j, & \text{if}\, \gamma_j > \bar{\gamma} + c\sigma \\ 0, & \text{otherwise} \end{cases},$$

where $\bar{\gamma}$ and σ is the mean and standard deviation calculated from $\gamma_j, j = 1, \ldots L$ and c is a prespecified parameter. The power of the test later should be dependent on the choice of c, which is usually set 1.28 or 1.64.

(2) Discordant Sibpair Design

For discordant sibpairs, assume the first sib is always chosen to be affected and the second is always unaffected, and there are N_{sib} discordant sibpairs. The weight of allele *A* at the *j*th SNP should be proportional to

$$\frac{P\left(sibs\ 1\ are\ affected\ and\ sib\ 2\ is\ not\ |\ (g_{i1j}, g_{i2j}) = (A\sim, aa)\right)}{p(sibs\ 1\ are\ affected\ and\ sib\ 2\ is\ not)}$$
$$= \frac{P\left((g_{i1j}, g_{i2j}) = (A\sim, aa)\ |\ sibs\ 1\ are\ affected\ and\ sib\ 2\ is\ not\right)}{\phi_3},$$

where $\phi_3 = P\left((g_{i1j}, g_{i2j}) = (A\sim, aa)\right) = \frac{1}{4} p_j \left(1 - p_j\right)^2 \left(4 - p_j\right)$. For the *i*th discordant sibpair and the *j*th SNP, the genotype score is

$$\tilde{g}_{ij} = \begin{cases} \frac{1}{\phi_3}, & \text{when } (g_{i1j}, g_{i2j}) = (A\sim, aa) \\ 0, & \text{otherwise} \end{cases}.$$

In the same way as for affected sibpairs, the weights for discordant sibpairs can be defined.

8.3 Discussion

Although there is a heat debate about the hypotheses of CDCV and CDRV, the identification and characterization of the effects of rare variants on common disease will play central parts in the future genetic studies. The contribution of the rare variants to complex diseases has already been reported for type 2 diabetes (Bonnefond et al. 2012), and rare variants will undoubtedly uncover some missing "heritability." However, more robust and powerful statistical methods for analyzing rare variants are still needed. The statistical methods discussed here will still need to be evaluated in practice. It should not be doubted that a better understanding of the genetic architecture and the underlying biology of complex diseases will help us to develop more powerful statistical methods to detect disease variants.

References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

Agresti A. Categorical data analysis. New York: Wiley-Interscience; 2002.

Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.

Bonnefond A, Clément N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S, Hercberg S, Hadjadj S, Balkau B, Marre M, Lantieri O, Langenberg C, Bouatia-Naji N, The Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC), Charpentier G, Vaxillaire M, Rocheleau G, Wareham NJ, Sladek R, McCarthy MI, Dina C, Barroso I, Jockers R, Froguel P. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat Genet. 2012;44:297–301.

Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005;37(2):161–5.

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A. 2006;103(6):1810–15.

Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.

Feng T, Elston RC, Zhu X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). Genet Epidemiol. 2011;35(5):398–409.

Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010;42(7):558–60.

Gudbjartsson DF, Walters GB, Thorleifsson G, et al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008;40:609–15.

Guo W, Lin S. Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet Epidemiol. 2009;33(4):308–16.

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010;70(1):42–54.

Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Magi R, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010;42(11):949–60.

Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, all PN, Klemm AK, and Manolio TA. A catalog of published genome-wide association studies 2011. Available at: www.genome.gov/gwastudies. Accessed Sep 15, 2012.

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010;5(11):e13584. doi:10.1371/journal.pone.0013584.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet. 2008;40(5):592–9.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–8.

Lettre G, Jackson AU, Gieger C, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet. 2008;40:584–91.

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83(3):311–21.

Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet. 2010;87(5):728–35.

Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011;89:354–67.

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7:e1001322.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009;324(5925):387–9.

Neyman J, Scott E. On the use of c(α) optimal tests of composite hypotheses. Bull Int Stat Inst. 1966;41:477–97.

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86:832–8.

Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001;69(1):124–37.

Ramensky V, Bork P, Sunyaev S. Human nonsynonymous SNPs: server and survey. Nucleic Acids Res. 2002;30:3894–900.

Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40:575–83.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). Am J Hum Genet. 2011;89:82–93.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet. 2010;87(5):604–17.

Zelterman D, Chen C. Homogeneity tests against central-mixture alternatives. J Am Stat Assoc. 1988;83(401):179–82.

Zhu X, Fejerman L, Luke A, Adeyemo A, Cooper RS. Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. Hum Mol Genet. 2005;14(5):639–43.

Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol. 2010;34(2):171–87.

Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012;109(4):1193–8.

Chapter 9
Gene Duplication and Functional Consequences

Xun Gu, Yangyun Zou, and Zhixi Su

Abstract Due to genome-wide or local chromosome duplication events, in almost all organisms, many genes are represented as several paralogs (duplicate genes) in the genome with related but distinct functions (gene families). After accumulating mutations in coding sequences and regulatory regions, duplicate genes may have more opportunity to diversify in protein function and/or regulatory network, leading to an increase of organismal complexity. With the explosive increase of high-throughput data, however, it has been realized that gene duplication is a much more complicated process, and the underlying mechanisms may vary considerably in different organisms. In this chapter, we shall focus on a few important issues related to copy-number variation, gene essentiality, transcriptional, and epigenetic divergence between duplicate genes.

Keywords Gene duplication • Copy-number variation • Gene essentiality • Protein function • Gene regulation

9.1 Introduction

Many organisms have undergone genome-wide or local chromosome duplication events during their evolution (Ohno 1970). As a result, many genes are represented as several paralogs in the genome with related but distinct functions (gene families).

X. Gu (✉)
Department of Genetics, Development and Cell Biology-Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University,
Ames, IA 50011, USA
e-mail: xungufudan@gmail.com

Y. Zou • Z. Su
School of Life Sciences, Center for Evolutionary Biology, Fudan University,
Shanghai 200433, China
e-mail: zxsu@fudan.edu.cn; yyzou@fudan.edu.cn

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_9, © Springer Science+Business Media Dordrecht 2012

Gene duplication and subsequent functional divergence have been universally regarded as an important means to provide raw genetic materials for evolution; see Innan and Kondrashov (2010) for a recent review. After accumulating mutations in coding sequences and regulatory regions, partly because of redundancy, duplicate genes may have more opportunity to diversify in protein function and/or regulatory network, leading to an increase of organismal complexity. The conceptual framework includes three main evolutionary fates of a newly born duplicate copy, that is, neofunctionalization (acquisition of new functions) (Ohno 1970), subfunctionalization (partitioning of ancestral functions) (Force et al. 1999; Lynch and Force 2000), or nonfunctionalization (being a pseudogene), as well as many combinations of above (Innan and Kondrashov 2010). With the explosive increase of high-throughput data, however, we have realized that gene duplication is a much more complicated process. Moreover, the underlying mechanisms may vary considerably in different organisms. Because of the complexity of this problem, in this chapter, we shall focus on a few issues to which our research group has made some contributions.

9.2 CNV and Segmental Duplications

Copy-number variation (CNV) is a new type of genetic variation. Technically, CNVs are defined as deletions or duplications greater than 1 kb in size. It was recently found that CNVs are abundant in the human genome; roughly, between any two individual humans, the number of base-pair differences due to CNVs is over 100-fold higher than that of SNPs (Zhang et al. 2009a; Conrad et al. 2010). Moreover, many diseases are associated with CNVs in the genome, such as autism and schizophrenia (Sebat et al. 2007; Lupski 2007; McCarthy et al. 2009; Pinto et al. 2010).

Study of tandem or segmental duplications (SD) has been advanced considerably by the discovery of CNVs that are pervasive in virtually all eukaryotes. The mutational mechanisms giving rise to CNVs in the population may underlie gene duplication and gene loss. In simple terms, a gene within the CNV region can be fixed and maintained in the population as a functional duplicate gene. Conversely, a duplicated copy could disappear in the population or remain in the genome as the relic of a pseudogene (Korbel et al. 2008). Hence, CNV and SD are the two facets of the same genomic dynamics at the population level and at the species level, respectively.

It is evident from genome-wide surveys that CNVs exhibit a highly nonuniform distribution along chromosomes (Nguyen et al. 2006; Cooper et al. 2007; Campbell et al. 2011). Similarly, the fixed segmental duplicated genes in the genome also show a large variation. For example, the mouse and opossum have more than 1,000 olfactory receptor genes but the human only has 387 olfactory receptor genes (Niimura and Nei 2007). Thus, one of most interesting issues about the relationship between CNVs and segmental duplicates is to what extent the CNVs' distribution and the following fixation process toward SDs are driven by the neutral process versus positive selection.

The nonrandom distribution of CNVs in the genome may have three potential causes. First, it may be due to technological biases in the detection of CNVs, but

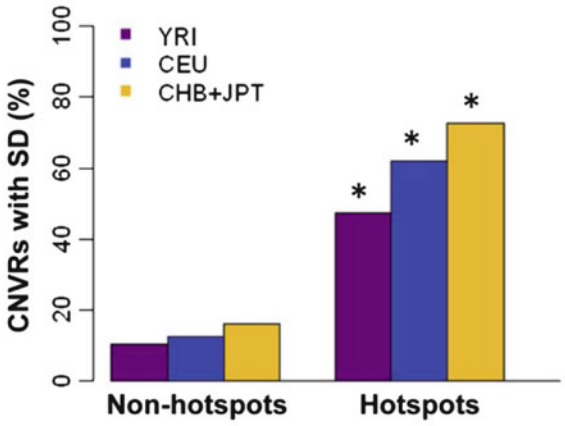

Fig. 9.1 The percentage of copy-number variation regions (*CNVRs*) overlapping with SDs (Figure adapted from Fu et al. 2010)

several analyses indicated that this explanation is unlikely to be responsible for the trend (Zhang et al. 2009a; Girirajan et al. 2011). Second, mutational hot spots, that is, locus-specific differences in the rate at which CNVs are formed, may cause this disparity. Finally, the third cause is natural selection acting differentially on different CNV events. It should be noticed that discriminating the remaining two potential causes (mutation or selection) is not straightforward. Though mutation bias, fixation by natural selection or random drifts have been studied extensively for SNPs, they have been much less studied for CNVs (Korbel et al. 2008).

A number of studies have shown that in the human and mouse, the mutation rate of CNV varies dramatically in different regions of each chromosome, resulting in a highly nonuniform distribution of CNVs along a chromosome (Korbel et al. 2008; Zhang et al. 2009a; Fu et al. 2010; Sudmant et al. 2010). For instance, Fu et al. (2010) conducted a genome-wide population genetic analysis and estimated that the mutation rate of CNVs varies up to 10^3-fold. Substantial evidence has shown that mutational mechanisms giving rise to copy-number mutations are strongly correlated with local genomic architecture, making particular regions systematically more prone to such mutations. For example, recent studies clearly showed that regions flanked by SDs of high sequence similarity are much more likely to harbor copy-number variation than other genomic sites, probably as a result of nonallelic homologous recombination (Sharp et al. 2005; Redon et al. 2006; Cooper et al. 2007). Fu et al. (2010) found that about 60% of CNV hot spots overlap with SDs, whereas only <20% of non-hot-spot CNV regions (CNVRs) involve SDs; $p = 7.63 \times 10^{-15}$ (Fisher's exact test) for YRI, $p = 4.39 \times 10^{-12}$ for CEU, and $p = 7.62 \times 10^{-9}$ for CHB + JPT (Fig. 9.1). Hence, subtelomeric regions represent hot spots for interchromosomal recombination (Linardopoulou et al. 2005) and segmental duplication of genomic sequence (Derti et al. 2006), as well as an enrichment of CNVs in subtelomeric regions (Redon et al. 2006). In addition, breakage or fusion of chromosomes during mammalian genome evolution may have influenced the rate of duplication (and loss) of gene families across species.

Fixation of duplicated genes in the population can be driven by natural selection or only by random effects. Nei et al. (2008) argued that, as long as the number of gene copies is within the upper and lower limits determined by the physiological requirements of the organism, intraspecific CNV variation is almost neutral. An observation used as strong evidence favoring the neutral CNV hypothesis is that such mutations can exist with weak or no phenotypic consequences. Indeed, the existence of CNV variations in a large sample of "normal" individuals indicates that many such mutations confer minimal to no phenotypic consequence within humans. At the very least, such copy-number variants do not have substantial deleterious effects. On the other hand, evolutionary studies of some large gene families as well as related pseudogenes support this hypothesis. For instance, Nozawa et al. (2007) reported that no significant difference exists in the amount of CNVs between functional and nonfunctional (pseudogene) sensory receptor genes, a gene family that has been found particularly prone to the structural variation (Redon et al. 2006; Korbel et al. 2007).

The effects of purifying selection imposed on copy-number variations are particularly visible for deletions. For instance, recent studies have revealed that protein-coding genes, and also other genomic elements including highly conserved noncoding regions, tend to be depleted among CNVs (Derti et al. 2006; Redon et al. 2006; Korbel et al. 2007). Cooper et al. (2007) invoked purifying selection to explain the enrichment in gene-poor regions for CNVs not overlapping SDs, because such variants would have a lower likelihood of disrupting protein-coding sequence than mutations in gene-rich regions. Furthermore, CNVs more often involve genes encoding proteins located at the periphery of the interaction network. This suggests that duplication of highly connected proteins can be deleterious, probably due to stoichiometric constraints to avoid abnormal phenotypes (Korbel et al. 2008). In several *Drosophila* populations, Dopman and Hartl (2007) have shown that genes encoding proteins with more protein–protein interactions are more likely to be underrepresented in CNVs, consistent with the claim that purifying selection on dosage-sensitive genes results in the removal of extra gene copies that may cause dosage imbalance.

In addition to purifying selection, positive selection has been implicated in shaping the distribution of CNVs and duplicated genes in the genome. Functional biases in the genes that are associated with CNVs may provide an indication of adaptive selection. The genes associated with environmentally responsive functions such as sensory perception and immunity tend to be affected by CNVs, while genes related to fundamental cellular processes are underrepresented (Feuk et al. 2006; Nguyen et al. 2006; Redon et al. 2006; Cooper et al. 2007; Korbel et al. 2008). By computational analysis, Jiang et al. (2007) reported some evidence for positive selection in hot spots of recently formed segmental duplications in humans, as these hot spots are presumably subject to recurrent *de novo* gene duplications. A variety of studies have found signs of positive selection at the level of amino acid replacement for recently duplicated genes in human and other species, such as morpheus69 and RanBP2 (Ciccarelli et al. 2005), or DUF1220 (Popesco et al. 2006) families.

Zhang et al. (2011) identified 1,828 young primate-specific genes in humans, most of which showed a rapid protein sequence evolution and are upregulated in the neocortex, the evolutionarily newest part of the human brain. As the timing of the emergence of these young genes (probably related to CNVs) was coincident with the evolutionary period during which the neocortex was expanding, they suggested a role of potential positive selection on the evolution of these genes. A recent study in the salivary amylase protein Amy1 showed that AMY1 gene copy number in human populations likely underlies diet-related positive selection pressures (Perry et al. 2007). However, the genetic biases of CNVs could be alternatively explained by the reduced efficiency of purifying selection in eliminating deleterious changes in humans (Nguyen et al. 2008). In some cases of CNVs spanning more than one gene, the positive effect of gene duplication or loss may balance or overshadow the potentially negative impact of protein dosage imbalances and consequently may drive the fixation of CNVs in particular regions of the genome.

9.3 Rate of Segmental Duplication and Rate of CNV

Fu et al. (2010) conducted a genome-wide population genetic analysis and estimated that the mutation rate of CNVs varies up to 10^3-fold, with the average roughly around 10^{-5} per locus per generation (excluding two extremes in both sides). Assuming that the generation of primates is roughly 15 years, the mutation rate of CNVs can be transformed as $10^{-5}/15 \approx 6.7 \times 10^{-7}$ per locus per year = 0.67 per locus per million years. That is quite a high mutation rate, compared with other mutation types such as SNPs. Note this estimate approximates the genome-wide average. For mutation hot spots of CNVs ($>10^{-3}$ per locus per generation), the turnover time for a single mutation occurring in a locus (gene) could be as short as 15,000 years. Since fixation of CNVs in the population is the origin of duplication genes, with respect to population genetics, it would be interesting to compare the mutation rate of CNVs and the evolutionary rate of new duplicate genes.

Bailey et al. (2002) estimated that at least 5% of the human genome consists of segmental duplications. They calculated that the span of the segmental duplications in the human genome ranges from ten to hundred of kilobases, with >90% identity at the nucleotide level between ancestral and duplicate segments. On the basis of neutral expectation of sequence divergence (molecular clock), this corresponds to duplications that have emerged over the past ~40 million years of human evolution. If these estimates are largely correct, one can obtain a rough estimate for the evolutionary rate of segmental duplications. Assume that the total number of genes in the ancestral primate genome in 40 million years (Myr) ago is ~25,000. Thus, the initial rate of segmental duplication can be roughly estimated as $v = 5\% \times 25000/40 \sim 31$ genes/Myr/genome. That is, on average, there are about 31 genes that are duplicated per million years in the genome.

On the other hand, it has been estimated that the retention frequency of duplicates could be from $f = 13\%$ (yeast) to 24% (*Arabidopsis*). If the fate of duplicates in human is similar to yeast or *Arabidopsis*, we find the emergence of new paralogous (functional) genes with the rate of $\lambda = f \times v \sim (4.0\text{–}7.6)$ genes/Myr/genome. Interestingly, Gu et al. (2002) estimated that the (stationary) rate of small-scale gene duplications is 3.2–5 genes/Myr/genome, during the course of vertebrate evolution but before the mammalian radiation. As segmental duplications are the major contribution to small-scale duplication events, one may conclude that these two rough estimates are close. Actually they are more similar when the genome size difference is taken into account. Assume that a vertebrate genome is on average, ~15,000 genes. Then, the rate of small-scale duplication during primate lineage can be revised as $(3.2\text{–}5)/15{,}000 \sim (2.1\text{–}3.3) \times 10^{-4}$ per gene per Myr. From Bailey et al.'s (2002) data, the rate is $(4.0\text{–}7.6)/25{,}000 \sim (1.6\text{–}3.0) \times 10^{-4}$ per gene per Myr.

We thus conclude that the emergence rate of new genes via segmental duplications may be roughly constant during the course of vertebrate evolution to primates, with the magnitude of $(2 \sim 3) \times 10^{-4}$ per gene per Myr. Yet, this estimate is considerably lower (<1%) than the mutation rate of CNVs, which is 0.67 per locus per million years. It seems that, in spite of abundant copy-number variation in the genome, only a very small portion of CNVs can be fixed in the population. Our tentative analysis supports the notion that the majority of CNVs are neutral or nearly neutral for a given range of copy number. A small portion of CNVs are deleterious or adaptive because of the dosage effects. Of course, more data and accurate analyses are needed for having a deep understanding of the relationship between copy-number variation and segmental duplication.

9.4 Essentiality (or Dispensability) in Duplicate Genes

Functional compensation of duplicate (paralogous) genes has been thought to play an important role in genetic robustness (Winzeler et al. 1999; Gu 2003; Gu et al. 2003; Kamath et al. 2003; Conant and Wagner 2004; Guan et al. 2007; Dean et al. 2008). Indeed, existence of a close paralog in the same genome could result in null mutations of the gene with little effect on the organismal fitness (nonessential gene), as observed in diverse organisms such as yeast, nematode, and Arabidopsis (Gu et al. 2003; Conant and Wagner 2004; Guan et al. 2007; Dean et al. 2008; Hanada et al. 2009, 2010; Gu 2010). In addition, large-scale double-knockout experiments in yeast and Arabidopsis demonstrate that the genetic redundancy is not just a transient consequence of gene duplication, but is often a long-term retained evolutionary stable state maintained by natural selection (Vavouri et al. 2008; Hanada et al. 2009; Li et al. 2010; VanderSluis et al. 2010; van Wageningen et al. 2010).

However, the role and magnitude of the duplicate genes contributing to genetic robustness in mammals remain controversial (Hsiao and Vitkup 2008; Liang and Li 2007; Liao and Zhang 2007; Su and Gu 2008; Wang and Zhang 2009; Makino et al. 2009; Liang and Li 2009; Qian et al. 2010). Two studies on mouse knockout phenotypes (Liang and Li 2007; Liao and Zhang 2007) observed that the proportion of

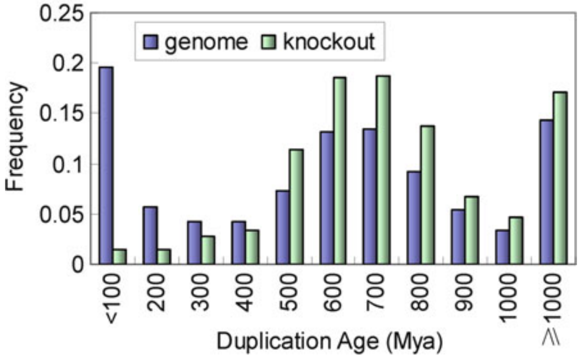

Fig. 9.2 Duplication age distribution of mouse genome set (*blue bars*) and knockout gene set (*green*)

essential genes (P_E) is similar between duplicate genes and singletons in mouse, sharply contrasted with those well-known findings that removing a duplicate gene usually generates less deleterious phenotypes than removing a singleton gene (Gu et al. 2003; Conant and Wagner 2004; Guan et al. 2007; Dean et al. 2008). Although recent reports pointed out the strong bias of mouse knockout data toward developmental genes, whole genome duplication-derived genes, or ancient duplicated genes, after correcting for these confounding factors, the relationship between phenotypic effect and gene duplication in the mouse still appears to be much weaker than those in the yeast and nematode (Su and Gu 2008; Makino et al. 2009; Liang and Li 2009). Moreover, the effect of genetic buffering (measured by P_E of singletons) is correlated with the protein sequence conservation as well as the protein–protein interactivity, making the efforts to seek the cause–effect relationship more complicated.

The effect of gene duplications on genetic robustness depends on the distribution of young duplicate genes in the current genome. Therefore, its impact varies among species, mainly because each species has its unique age distribution of gene duplications. For instance, due to recent polyploidizations, duplicate genes may dominate the genetic robustness in plant genomes (Wendel 2000). Indeed, we (Su and Gu 2008) found that the effect of duplicate genes on mouse genetic robustness is duplication age dependent. The histogram in Fig. 9.2 clearly shows that mouse knockout experiments have been designed to avoid recently duplicated genes. For example, only 1.4% of duplicated genes in the knockout set were dated within 100 mya (around or after the mammalian radiation), compared to 19.6% in the mouse genome set. Consequently, the ages of duplicate genes in the mouse knockout dataset are typically around 500–700 mya (in early vertebrates), with a long tail toward even more ancient ones (>1,000 mya). In other words, the sampling bias toward ancient duplicates in the currently available mouse knockout target genes has been nontrivial. Recently, gene duplications, those duplicated around the mammalian radiation or in the rodent lineage, are expected to have significant contributions to the genetic robustness in the current mouse genome. While these young duplicates were considerably underrepresented in the mouse knockout dataset, the observed proportion of essential duplicate genes is upwardly biased toward the value of singletons.

We reanalyzed the updated mouse knockout phenotype data in Mouse Genome Database (MGD) (Eppig et al. 2005) (unpublished). Consistent with previous studies (Liang and Li 2007; Liao and Zhang 2007; Su and Gu 2008), the updated mouse knockout dataset shows no statistical difference of P_E between singletons and duplicates (47% vs. 46.3%; $P>0.05$), which holds after ruling out the potential confounding effect from coding sequence conservation, protein–protein connectivity, functional bias, or the bias of duplicates generated by whole genome duplication (WGD). After using a simple bias-correcting procedure (Su and Gu 2008) to calculate a bias-corrected P_E, we predicted that $P_E=41.7\%$ for all duplicate genes, which is impressive, compared to $P_E=46.3\%$ observed in sample duplicates and $P_E=47\%$ in sample singletons (Su and Gu 2008). However, we emphasize here that even after taking this sampling bias into consideration, the difference between P_E for singletons and P_E for duplicates at the mouse genome level remains small.

We call this controversy *the mouse knockout duplicate puzzle*. Resolving this issue may have a significant impact on biomedical sciences since knockout mice have been widely used as animal models of human diseases. There is no disagreement that some of these ancient duplicates may have undergone substantial functional divergence to have lost the capacity of functional compensation. Consequently, the contribution of functional compensation by young duplicates has been canceled by the contribution of higher intrinsic importance of ancient duplicates (Liang and Li 2009). Yet, the central issue for the underlying mechanism remains unsolved. We speculate that rapid increase of organismal complexity (as measured by the number of cell types in the early stage) may play a crucial role on resolving the mouse knockout duplicate puzzle.

9.5 Transcriptional Regulation Divergence Following Gene Duplication

Gene duplication is an evolutionary force for increased diversity and complexity of gene regulation and expression (Gu et al. 2004; Teichmann and Babu 2004), which facilitates an organism's adaptation to environmental changes or other biological competitions. It is commonly accepted that expression divergence between duplicate genes is an importantly first step of functional divergence after the gene duplication. A considerable amount of research studies has been published in attempt to unveil the underlying mechanism of expression divergence between duplicates. In the following we briefly discuss three types of regulation modes: *cis*-regulation, *trans*-regulation and epigenetic regulation.

Cis-regulatory divergence: As the most direct effect on gene expression regulation, how the *cis*-regulatory control diverges after gene duplication has been the hot topic in this field. It has been reported that duplicate genes of many organisms have experienced rapid *cis*-regulatory divergence and thus specialization of gene regulatory control, including the yeasts (Papp et al. 2003; Tirosh and Barkai 2007; Singh and Hannenhalli 2010), insects (Nielsen et al. 2010; Datta et al. 2011), plants (Lockton and Gaut 2005; Chen et al. 2010), and vertebrates (Bird et al. 2007; Woolfe and Elgar 2007; Kostka et al. 2010; Lee et al. 2010; Nowick et al. 2010). Rapid *cis*-regulatory divergence and specialization may be driven by the accelerated evolution in the noncoding sequence, probably due to the relaxed selective constraints in redundant duplicates (Bird et al. 2007). For example, Woolfe and Elgar (2007) analyzed seven pairs of teleost-specific paralogs involved in early vertebrate development and observed a pattern of *cis*-element retention and loss between Fugu paralogs, implying possible regulatory subfunctionalization. Subfunction in expression domains was also found in insects and plants (Lockton and Gaut 2005; Nielsen et al. 2010). Papp et al. (2003) argued that, except for the degenerative complementation, positive selection may occur on the *cis*-regulatory motif after yeast gene duplication, contributing to the regulatory neofunctionalization between yeast duplicates. In addition, fast evolution of exonic splicing enhancers and silencers shortly after the gene duplication provided another level of regulatory differentiation in duplicate genes (Zhang et al. 2009b).

Trans-regulatory divergence: In spite of the positive correlation between *cis*-regulatory motif divergence and the expression divergence of duplicate genes that is biologically intuitive (Castillo-Davis et al. 2004; Zhang et al. 2004), only a limited amount (about 2–3%) of expression variation can be explained by *cis*-motif divergence (Zhang et al. 2004). This observation suggested that other *trans*-acting factors may play some important roles in influencing the pattern of expression divergence. Using yeast regulatory interaction data (transcription factor (TF) target gene) by the ChIP technology (Lee et al. 2002) and yeast microarrays, Gu et al. (2005) estimated that, after the gene duplication, the evolutionary rate of regulatory interactions is, on average, about one order of magnitude (tenfold) faster in the young duplicates than that in the ancient duplicates, indicating a rapid evolution of gene expression shortly after the gene duplication. To provide an overview of the full landscape of regulatory network evolution, a novel strategy termed "genetical genomics" was proposed to address quantitative variation in gene expression, called expression quantitative trait loci (eQTL); see Jansen and Nap (2001) for a recent review. By this approach, Zou et al. (2009b) applied the yeast genome-wide *trans*-acting eQTL data (*trans*-genetic variation responsible for the gene expression variation) to investigate the evolutionary pattern of the genetic regulatory system between duplicate genes. The main results of this study are (1) the divergence of *trans*-acting eQTLs between duplicate pairs increases with evolutionary time, using the distance of synonymous substitutions or nonsynonymous substitutions between duplicates as a proxy (Fig. 9.3), and (2) *trans*-acting eQTL divergence can explain about 21% of the variation in expression divergence between young duplicate pairs (using the cutoff $K_S < 2.0$); when the TF-target interactions are combined, the proportion of explainable variation can be up to 27%.

Epigenetic divergence: An increasingly accepted evolutionary force for expression divergence between duplicate genes is differential epigenetic control. Different epigenetic stage/tissue-complementary silencing patterns by DNA methylation

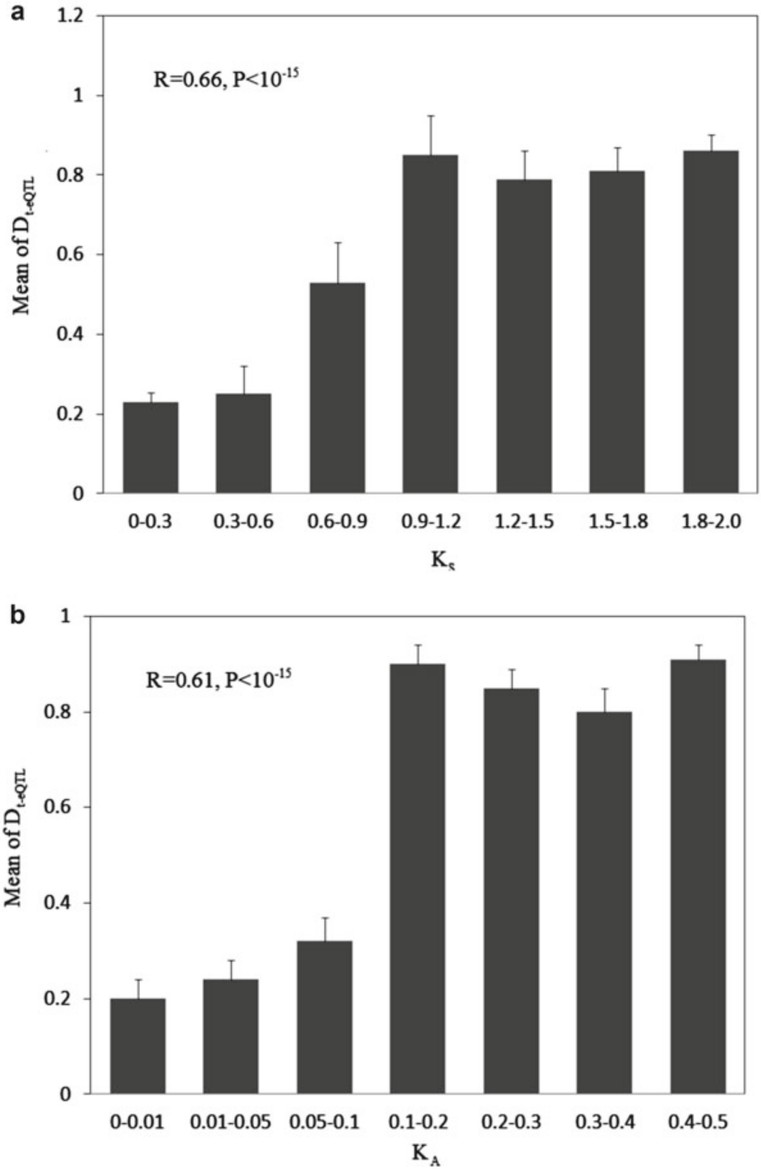

Fig. 9.3 Divergence of *trans*-acting eQTLs between duplicate pairs ($D_{t\text{-}eQTL}$) increases with synonymous distance K_S (panel *A*) or nonsynonymous distance K_A (panel *B*) between duplicate pairs. Error bar indicates standard error

between duplicate pairs may favor recently duplicated genes to survive (Rodin and Riggs 2003). It was suggested that many newly born duplicates will be translocated to ectopic chromosome locations, often on different chromosomes, to escape the fate of pseudogenization. As a result, this may create a different chromatin environment and thus lead to different epigenetic regulation between two duplicate copies (Rodin and Parkhomchuk 2004; Rodin et al. 2005). MicroRNA (miRNA) is also a potential epigenetic factor affecting gene regulation evolution (Huang and Gu 2011), through sophisticated regulation and evolution on the 3′ UTR where most miRNA target sites are located. After investigating the miRNA-mediated transcriptional regulation between human and mouse duplicate genes, Li et al. (2008) found that miRNA targets are significantly overrepresented in duplicate genes and shared miRNA regulators between them decrease with evolutionary time. Moreover, ancient duplicates seem more likely to be regulated by miRNAs, probably because of the acquirement of miRNA regulation over time since gene duplication or underpresentation of young human/mouse duplicates involved in miRNA regulation.

9.6 Gene Duplication and Environmental Adaptation

An evolving system is the process of organisms constantly adapting themselves to the changing environment through natural selection. High variable environmental circumstances require that organisms have multiple and sensitive stress sensing and response mechanisms which demand constant innovation. Gene duplication may provide new genetic materials for adapting to the changing environment. Many investigations have observed that duplicate genes are associated with biological processes interacting with the external environments in both prokaryotes (Sanchez-Perez et al. 2008; Bratlie et al. 2010; Chia and Goldenfeld 2011) and eukaryotes (Moore and Purugganan 2005; Rizzon et al. 2006; Hanada et al. 2008; Ames et al. 2010). Several examples are as follows: (1) Ames et al. (2010) demonstrated that yeast lineage-specific duplicate genes and strain-specific duplicates (or CNV) are abundant, with detectable bias in specific functions correlated with the environment from which they were isolated. (2) Chen et al. (2008) and Podrabsky (2009) provided a detailed case study illustrating the important role of gene duplication in the adaptation to subfreezing temperatures in the Antarctic notothenioid fish. (3) Yamanaka et al. (1998) studied the CspA family, the major cold-shock protein in *E. coli.* The authors observed nine members of the family from *cspA* to *cspl* generated by a series of gene duplications, many of which have particular roles of responding to different environmental stresses, such as *cspA*, *cspB,* and *cspG* responsible for cold-shock stress and *cspD* for nutritional deprivation. And (4) other well-known examples include the expansion of the olfactory receptor gene family in mammalian and adaptive immunity-related genes such as major histocompatibility complex (MHC), T-cell receptors (TCR), and immunoglobulins (Ig) in vertebrate genomes by gene duplications (Firestein 2001; Azumi et al. 2003; Niimura and Nei 2003). Increased sensory adaptation to different odorant molecules in the environment influences mammalian behavior considerably influencing food-seeking, mate and offspring-identifying, as well as danger-escaping. Meanwhile, genetic and somatic diversity of immune-related genes help organisms to prevail in an evolutionary “arms race” with pathogens.

Table 9.1 Ancestral TATA box state reconstruction for *S. cerevisiae* gene families by parsimony principle

Two paralogs		At least three paralogs	
Ancestral TATA box state	Number of families	Ancestral TATA box state	Number of families
TATA (+)	49	TATA (+)	27
TATA (−)	368	TATA (−)	161
Ambiguity	125	Ambiguity	37

Table 9.2 Events of TATA box switches during the yeast gene family evolution according to different parsimony optimizations

Parsimony optimization	TATA box gains	TATA box losses	Binomial test
ACCTRAN	75	26	$P < 10^{-5}$
DELTRAN	78	23	$P < 10^{-7}$

In addition to providing new genetic materials, gene duplication and subsequent divergence can offer diversified regulatory system with certain regulatory elements for variable regulations, which allows organisms to flexibly deal with different external stimuli (Lopez-Maury et al. 2008). Zou et al. (2009a) explored the evolution of stress-regulated gene expression among duplicate genes in *Arabidopsis thaliana* after reconstructing the putative ancestral stress regulation patterns. They observed that duplicate genes experienced substantial changes (loss, gain, or switch) of stress responses, especially for the lost events. Interestingly, ancestral stress response partitioning was highly asymmetric between duplicate genes, as well as differential losses of DNA regulatory elements. Hence, (asymmetric) mutations in the regulatory element after the gene duplication help the organism to improve the capacity of flexible responsiveness to altered environmental stresses.

We (Zou et al. 2011) recently provided direct evidence which supports the notion that gene duplication may contribute to environmental adaptation by providing new stress sensing and regulatory response mechanisms through the gain of TATA box, a core promoter element in eukaryote. TATA box motif appears to play a unique role in stress-related, multi-stimulus response genes, as shown to be associated with variably expressed genes (Basehoar et al. 2004; Walther et al. 2007). TATA box is significantly enriched in duplicate genes compared with singletons in the human, worm, *Arabidopsis,* and yeast. We further conducted extensive genomic analyses to investigate the evolution of TATA box among over 700 yeast gene family phylogenies. After reconstructing the ancestral TATA box states (presence or absence) that were usually TATA box absent (Table 9.1), we (Zou et al. 2011) found a significantly higher number of TATA box gains than losses that occurred after yeast gene duplications; the overall gain/loss ratio is about 3–4 (Table 9.2). Our result suggests that the enrichment of the TATA box in yeast duplicate genes may be the consequence of consecutive gains of new TATA boxes since gene duplication. These TATA-gain duplicate genes, on average, have experienced greater expression divergence than their closely related TATA-less genes (genes without TATA box in the promoter) duplicate partners, only under environmental stress conditions. Under normal physiological conditions, they have similar expression divergence. Besides TATA-gain duplicates, that is, acquiring a new TATA box after gene duplication, stress-associated functional categories are enriched, such as transport, cell membrane, and extracellular process. Putting everything together, Zou et al. (2011) concluded that gain of the TATA box (a stress-sensitive regulatory motif) after gene duplication may be an important mechanism for organisms to adapt to drastically changing environments via more flexible and sensitive expression regulation program which also leads to the preservation of such duplicate genes in the genome.

9.7 Concluding Remarks: Toward a New View of Gene Duplication?

Almost in all eukaryotes, gene duplication is the major mechanism to provide new genetic material resources for increasing the diversity and complexity of protein function and gene regulation, facilitating adaptation to environmental change or other biological competition. However, the underlying mechanisms to achieve this goal may differ considerably among different organisms, alongside differences in the ancient environments when the duplications occurred. Hence, we question some recent efforts aimed at the establishment of a universal model for duplicate preservation and divergence that can be conceptually applied to all organisms.

References

Ames RM, Rash BM, Hentges KE, Robertson DL, Delneri D, Lovell SC. Gene duplication and environmental adaptation within yeast populations. Genome Biol Evol. 2010;2:591–601.

Azumi K, Santis R, Tomaso A, Rigoutsos I, Yoshizaki F, Pinto M, Marino R, Shida K, Ikeda M, Ikeda M, Arai M, Inoue Y, Shimizu T, Satoh N, Rokhsar D, Pasquier L, Kasahara M, Satake M, Nonaka M. Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "waiting for Godot". Immunogenetics. 2003;55:570–81.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. Science. 2002;297:1003–7.

Basehoar AD, Zanton SJ, Pugh BF. Identification and distinct regulation of yeast TATA box-containing genes. Cell. 2004;116:699–709.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, MillerO W, Hurles ME, Dermitzakis ET. Fast-evolving noncoding sequences in the human genome. Genome Biol. 2007;8:R118.

Bratlie M, Johansen J, Sherman B, Huang DW, Lempicki R, Drablos F. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics. 2010;11:588.

Campbell CD, Sampas N, Tsalenko A, Sudmant PH, Kidd JM, Malig M, Vu TH, Vives L, Tsang P, Bruhn L, Eichler EE. Population-genetic properties of differentiated human copy-number polymorphisms. Am J Hum Genet. 2011;88:317–32.

Castillo-Davis CI, Hartl DL, Achaz G. *Cis*-regulatory and protein evolution in orthologous and duplicate genes. Genome Res. 2004;14:1530–6.

Chen Z, Cheng C-HC, Zhang J, Cao L, Chen L, Zhou L, Jin Y, Ye H, Deng C, Dai Z, Xu Q, Hu P, Sun S, Shen Y, Chen L. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc Natl Acad Sci U S A. 2008;105:12944–9.

Chen KN, Zhang YB, Tang TA, Shi SH. *Cis*-regulatory change and expression divergence between duplicate genes formed by genome duplication of *Arabidopsis thaliana*. Chin Sci Bull. 2010;55:2359–65.

Chia N, Goldenfeld N. Dynamics of gene duplication and transposons in microbial genomes following a sudden environmental change. Phys Rev E Stat Nonlin Soft Matter Phys. 2011;83:021906.

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. Complex genomic rearrangements lead to novel primate gene function. Genome Res. 2005;15:343–51.

Conant GC, Wagner A. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. Proc Biol Sci. 2004;271:89–96.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464:704–12.

Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39:S22–9.

Datta RR, Cruickshank T, Kumar JP. Differential selection within the Drosophila retinal determination network and evidence for functional divergence between paralog pairs. Evol Dev. 2011;13:58–71.

Dean EJ, Davis JC, Davis RW, Petrov DA. Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet. 2008;4:e1000113.

Derti A, Roth FP, Church GM, Wu CT. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet. 2006;38:1216–20.

Dopman EB, Hartl DL. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2007;104:19920–5.

Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, The Mouse Genome Database Group. The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. Nucl Acids Res. 2005;33:D471–5.

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97.

Firestein S. How the olfactory system makes sense of scents. Nature. 2001;413:211–18.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151:1531–45.

Fu W, Zhang F, Wang Y, Gu X, Jin L. Identification of copy number variation hotspots in human populations. Am J Hum Genet. 2010;87:494–504.

Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011;45:203–26.

Gu X. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. 2003;19:354–6.

Gu X. Statistical theory and methods for evolutionary genomics. New York: Oxford University Press; 2010.

Gu X, Wang Y, Gu J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet. 2002;31:205–9.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. Role of duplicate genes in genetic robustness against null mutations. Nature. 2003;421:63–6.

Gu Z, Rifkin SA, White KP, Li W-H. Duplicate genes increase gene expression diversity within and between species. Nat Genet. 2004;36:577–9.

Gu X, Zhang ZQ, Huang W. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci U S A. 2005;102:707–12.

Guan Y, Dunham MJ, Troyanskaya OG. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. Genetics. 2007;175:933–43.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 2008;148:993–1003.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. Genome Biol Evol. 2009;1:409–14.

Hanada K, Sawada Y, Kuromori T, Klausnitzer R, Saito K, Toyoda T, Shinozaki K, Li WH, Hirai MY. Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. Mol Biol Evol. 2010;28:377–82.

Hsiao TL, Vitkup D. Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet. 2008;4:e1000014.

Huang Y, Gu X. A study of the evolution of human microRNAs by their apparent repression effectiveness on target genes. PLoS One. 2011;6:e25034.

Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 2010;11:97–108.

Jansen RC, Nap J-P. Genetical genomics: the added value from segregation. Trends Genet. 2001;17:388–91.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet. 2007;39:1361–8.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature. 2003;421:231–7.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007;318:420–6.

Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr Opin Struct Biol. 2008;18:366–74.

Kostka D, Hahn MW, Pollard KS. Noncoding sequences near duplicated genes evolve rapidly. Genome Biol Evol. 2010;2:518–33.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science. 2002;298:799–804.

Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol. 2010;25:1205–15.

Li J, Musso G, Zhang Z. Preferential regulation of duplicated genes by microRNAs in mammals. Genome Biol. 2008;9:R132.

Li J, Yuan Z, Zhang Z. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet. 2010;6:e1001187.

Liang H, Li W-H. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 2007;23:375–8.

Liang H, Li WH. Functional compensation by duplicated genes in mouse. Trends Genet. 2009;25:441–2.

Liao B-Y, Zhang J. Mouse duplicate genes are as essential as singletons. Trends Genet. 2007;23:378–81.

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature. 2005;437:94–100.

Lockton S, Gaut BS. Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 2005;21:60–5.

Lopez-Maury L, Marguerat S, Bahler J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat Rev Genet. 2008;9:583–93.

Lupski JR. Genomic rearrangements and sporadic disease. Nat Genet. 2007;39:S43–7.

Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–73.

Makino T, Hokamp K, McLysaght A. The complex relationship of gene duplication and essentiality. Trends Genet. 2009;25:152–5.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimaki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nothen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J. Microduplications of 16p11.2 are associated with. Nat Genet. 2009;41:1223–7.

Moore RC, Purugganan MD. The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol. 2005;8:122–8.

Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 2008;9:951–63.

Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS Genet. 2006;2:e20.

Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res. 2008;18:1711–23.

Nielsen MG, Gadagkar SR, Gutzwiller L. Tubulin evolution in insects: gene duplication and subfunctionalization provide specialized isoforms in a functionally constrained gene family. BMC Evol Biol. 2010;10:113.

Niimura Y, Nei M. Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci U S A. 2003;100:12235–40.

Niimura Y, Nei M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One. 2007;2:e708.

Nowick K, Hamilton AT, Zhang HM, Stubbs L. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. Mol Biol Evol. 2010;27:2606–17.

Nozawa M, Kawahara Y, Nei M. Genomic drift and copy number variation of sensory receptor genes in humans. Proc Natl Acad Sci U S A. 2007;104:20421–6.

Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.

Papp B, Pal C, Hurst LD. Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet. 2003;19:417–22.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256–60.

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, Piven J, Ponting CP, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Sequeira AF, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stein O, Sykes N, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Van Engeland H, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Webber C, Weksberg R, Wing K, Wittemeyer K, Wood S, Wu J, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Devlin B, Ennis S, Gallagher L, Geschwind DH, Gill M, Haines JL, Hallmayer J, Miller J, Monaco AP, Nurnberger Jr JI, Paterson AD, Pericak-Vance MA, Schellenberg GD, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Scherer SW, Sutcliffe JS, Betancur C. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466:368–72.

Podrabsky JE. Gene duplication underlies cold adaptation in Antarctic fish. J Exp Biol. 2009;212:v–vi.

Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science. 2006;313:1304–7.

Qian W, Liao BY, Chang AY, Zhang J. Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. 2010;26:425–30.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature. 2006;444:444–54.

Rizzon C, Ponger L, Gaut BS. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol. 2006;2:e115.

Rodin SN, Parkhomchuk DV. Position-associated GC asymmetry of gene duplicates. J Mol Evol. 2004;59:372–84.

Rodin SN, Riggs A. Epigenetic silencing may aid evolution by gene duplication. J Mol Evol. 2003;56:718–29.

Rodin SN, Parkhomchuk DV, Rodin AS, Holmquist GP, Riggs AD. Repositioning-dependent fate of duplicate genes. DNA Cell Biol. 2005;24:529–42.

Sanchez-Perez G, Mira A, Nyiro G, Pasi L, Rodriguez-Valera F. Adapting to environmental changes using specialized paralogs. Trends Genet. 2008;24:154–8.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong association of de novo copy number mutations with autism. Science. 2007;316:445–9.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 2005;77:78–88.

Singh LN, Hannenhalli S. Correlated changes between regulatory cis elements and condition-specific expression in paralogous gene families. Nucleic Acids Res. 2010;38:738–49.

Su Z, Gu X. Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. J Mol Evol. 2008;67:705–9.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. Diversity of human copy number variation and multicopy genes. Science. 2010;330:641–6.

Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nat Genet. 2004;36:492–6.

Tirosh I, Barkai N. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol. 2007;8:R50.

van Wageningen S, Kemmeren P, Lijnzaad P, Margaritis T, Benschop JJ, de Castro IJ, van Leenen D, Groot Koerkamp MJ, Ko CW, Miles AJ, Brabers N, Brok MO, Lenstra TL, Fiedler D, Fokkens L, Aldecoa R, Apweiler E, Taliadouros V, Sameith K, van de Pasch LA, van Hooff SR, Bakker LV, Krogan NJ, Snel B, Holstege FC. Functional overlap and regulatory links shape genetic interactions between signaling pathways. Cell. 2010;143:991–1004.

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol Syst Biol. 2010;6:429.

Vavouri T, Semple JI, Lehner B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet. 2008;24:485–8.

Walther D, Brunnemann R, Selbig J. The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. PLoS Genet. 2007;3:e11.

Wang Z, Zhang J. Abundant indispensable redundancies in cellular metabolic networks. Genome Biol Evol. 2009;1:23–33.

Wendel JF. Genome evolution in polyploids. Plant Mol Biol. 2000;42:225–49.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science. 1999;285:901–6.

Woolfe A, Elgar G. Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biol. 2007;8:R53.

Yamanaka K, Fang L, Inouye M. The CspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. Mol Microbiol. 1998;27:247–55.

Zhang ZQ, Gu JY, Gu X. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? Trends Genet. 2004;20:403–7.

Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009a;10:451–81.

Zhang ZG, Zhou L, Wang P, Liu Y, Chen XF, Hu LDA, Kong XY. Divergence of exonic splicing elements after gene duplication and the impact on gene structures. Genome Biol. 2009b;10:R120.

Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 2011;9:e1001179.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet. 2009a;5:e1000581.

Zou Y, Su Z, Yang J, Zeng Y, Gu X. Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. J Exp Zool B Mol Dev Evol. 2009b;312B:722–33.

Zou Y, Huang W, Gu Z, Gu X. Predominant gain of promoter TATA box after gene duplication associated with stress responses. Mol Biol Evol. 2011;28:2893–904.

Chapter 10
From GWAS to Next-Generation Sequencing on Human Complex Diseases: The Implications for Translational Medicine and Therapeutics

Hai-De Qin, Alan Scott, Harold Z. Wang, and Yin Yao Shugart

Abstract Because the genetic architecture of human complex disease remains elusive, the clinical application of current genetic knowledge remains a formidable challenge. The field of translational genomics is rapidly progressing with the development of next-generation sequencing (NGS) technologies. NGS is poised to produce a sea change in clinical practice by providing new insights into our understanding of complex human diseases. In this review, we will discuss a broad range of issues on how NGS will transform current genetic research especially in the areas of cancer and mental health.

Keywords Next-generation sequencing • Translational medicine • Massive parallel sequencing • Human complex disease • Cancer • Mental disorder

H.-D. Qin
Unit of Statistical Genomics, Division of Intramural Research Programs,
National Institute of Mental Health (NIMH)/NIH, Building 35; Room 3A-1006,
Bethesda, MD, USA
e-mail: qinh2@mail.nih.gov

A. Scott
Department of Medicine, Johns Hopkins University, 600 North Wolfe Street,
Blalock Building, Room 1033, Baltimore, MD, USA

H.Z. Wang • Y.Y. Shugart (✉)
Unit of Statistical Genomics, Intramural Research, Program,
National Institute of Mental Health, 35 Convent Drive,
Room 3A1000, Bethesda, MD 20892-3719, USA
e-mail: wangh13@mail.nih.gov; kay1yao@mail.nih.gov

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1,
DOI 10.1007/978-94-007-5558-1_10, © Springer Science+Business Media Dordrecht 2012

10.1 Introduction

Cancer and mental disorders are two groups of disorders in which the contribution of genetic components to the risk of disease is substantial and in which multiple environmental factors are also involved. Mathematically, the pathogenesis of complex disease can be described as multiple random events occurring in multidimensional space. The pathogenesis process can undergo selection as the result of accumulating events during disease progression, thereby producing multiple phenotypes that can be recognized as a complex disease.

Worldwide, there were 12.6 million new incident cases of cancer in 2008 with 7.6 million deaths (Ferlay et al. 2010). Narrowing the distance between laboratory research and clinical applications is challenging but urgently needed. In the past several decades, scientists have made strong efforts to fill in this gap. Although some findings have led to the invention of drugs, the majority of patients with cancer would benefit mainly from the early screening program, surgery, and radiotherapy. Chemotherapy, for example, has little effect on cancer treatment for patients with late-stage metastatic diseases. As cancer progresses, cancer cells acquire multiple somatic mutations resulting in highly heterogeneous genetic patterns and the development of different phenotypes. Consequently, for any given chemotherapy regime, some cancer cells are likely to escape.

It has been recognized that poor diets, tobacco uses, virus infections, occupational exposures, and environmental pollutions count for over 65% of cancer causes (The American Cancer Society 2007). Most of the environmental risk factors have profound impacts on the genome, producing both somatic mutations and epigenetic changes. These environmental insults accumulate and are stored in the human genome. Genetic alteration is not only the hallmark of cancer but also the driving force, leading to the selection of clonal populations that results in progression, relapse, and metastasis. The simultaneous profiling of genomic alterations, including deletions, insertions, substitutions, and structural variations of patients with cancer, is technically feasible, and techniques are being developed to use unique genomic features of a patient's tumour (e.g. translocations) to follow progression and optimize treatment strategies (Leary et al. 2010).

Compared to cancer, neuropsychiatric disorders are usually not lethal, but they have profound negative effects on people's daily lives and may last the course of a lifetime. For example, some mental disorders associated with permanent brain dysfunction manifest their symptoms as early as childhood. About 30% of the population suffers from some form of mental disorder, and the prevalence has remained unchanged despite increased awareness and new treatments (Kessler and Wang 2008; Kessler et al. 2005; MacKenzie et al. 2005). It is estimated that only 30% of patients with mental disorders received treatment between 2001 and 2003 (Bartley et al. 2011; Chukwujama and Gormley 2011; Petersen 2011). Schizophrenia, bipolar, cognitive impairment, and autism are serious disorders that affect a large number of the population. For example, about 6 children per 10,000 are diagnosed with autism, and at least 18.7 per 10,000 are diagnosed with some form of pervasive developmental disorder. These numbers highlight the need for special services and education for a large group of children (Fombonne 1999).

To date, diagnosis and treatment of neuropsychiatric disease is still challenging due largely to the unclear mechanisms of aetiology. To diagnose a patient with cancer, oncologists use a variety of imaging tools such CT, MRI, and PET, and then confirm and further classify the tumours based on histology. Psychiatrists must evaluate a patient's mental status, personality, and cognitive function and sometimes use neurophysiologic measurements and sophisticated neuroimaging techniques (e.g. PET scan). As a result, a psychiatrist faces the dilemma of distinguishing a patient's disease from over 300 subtypes of mental illness (American Psychiatric Association 2000), and the resulting diagnosis largely relies on the experience of the psychiatrist. Psychiatrists also dispute the underlying basis for different mental illnesses, and due to misclassification or heterogeneity of the diseases, the therapeutics they select may not be the most effective. Moreover, therapy by antipsychotic drugs, antidepressants, mood stabilizers, and antianxiety agents, which are non-targeting drugs, might produce significant side effects depending on the disease or the patient's ability to metabolize a particular pharmaceutical.

Translational medicine, based on understanding genomic architecture, is expected to help overcome the diagnostic difficulties for both cancers and mental illnesses by illuminating common patterns of genetic change in specific subtypes of diseases. The invention of next-generation sequencing (NGS), including applications such as whole genome sequencing (WGS), whole exome sequencing (WES), RNA-Seq, Met-seq, and ChIP-Seq, has resulted in powerful new ways to interrogate the genome. NGS is already transforming genetic research and clinical practice by providing us with a much better resolution of underling genetic mechanisms for cancer, mental illnesses, and a spectrum of other diseases. Already, diagnoses are being made with greater precision, and treatments are being individualized based on a more thorough understanding of each patient. In this review, we focus on this rapidly progressing field, covering topics on the application of NGS to translational medicine in oncology and psychiatry.

10.2 Pre-NGS: Genome-Wide Association Study on Cancers and Mental Disorders

By using SNP genotyping arrays that target common variants, hundreds of genome-wide association studies (GWAS) have been published since 2007. A summary of these studies is available at www.genome.gov/gwastudies/. This approach was premised on the 'common diseases, common variants' hypothesis and was made possible by (1) technologies, largely by Affymetrix and Illumina, to rapidly and inexpensively genotype thousands to millions of single nucleotide polymorphisms (SNP) and (2) by the HapMap project that identified these variants by sequencing the genomes of samples collected from populations around the world. In this section we summarize some of the findings from GWAS studies in cancer and mental illness.

10.2.1 Using GWAS to Identify Risk Variants for Cancers

Over 30 (sub)types of tumours have been investigated by GWAS, including solid tumours, leukaemia, and lymphomas. For breast cancer, multiple GWAS studies have confirmed reproducible signals near *FGFR2*, the region between *TOX3* to *CHD9*, and near *TNP1* to *DIRC*. Twenty-three other genes significantly associated with breast cancer risk were also found (Table 10.1). For prostate cancer, GWAS studies have identified 46 loci that have shown association with cancer risks. Among them *CALM2P1-SOX9, HNF1B, FAM84B-SRRM1P1, CXorf67-NUDT11, LMTK2, LOC100289162, MSMB, PDLIM5, POU5F1B-MYC, SRRM1P1-POU5F1B, TPCN2-MYEOV,* and *VGLL3-CHMP2B* are confirmed by multiple GWAS studies (Table 10.1). Furthermore, *AGPHD1, BAT3, CHRNA3, CLPTM1L, TERT,* and *TP63* have been found to be significantly associated with lung cancer risk in multiple GWAS studies, and additional evidence suggest the same for at least 9 additional genes. For hepatocellular carcinoma, GWAS studies identified *HLA-DQB1–HLA-DQA2, HLA-S-MICA,* and *KIF1B*. Moreover, GWAS studies identified at least 14 such loci/genes for pancreatic cancer and 4 such genes for gastric cancer (Table 10.1).

10.2.2 GWAS Studies on Mental Disorders

Several mental illnesses have been investigated by GWAS, including bipolar disorder, schizophrenia, autistic disorder, Alzheimer's disease, and obsessive–compulsive disorder (OCD). The pioneering GWAS study in complex genetic disorders was conducted by the Wellcome Trust Case Control Consortium (WTCCC), which included 14,000 cases of seven common diseases and 3,000 shared controls. For bipolar disorder, the results suggested that the SNPs in the gene *CACNA1C* showed strong evidence of association (WTCCC 2007). That result was validated in a study of 4,387 cases and 6,209 controls by another group at the Massachusetts General Hospital (Ferreira et al. 2008) who, in addition to *CACNA1C*, identified a novel locus, *ANK3*. By 2011, at least 15 genes/loci were shown by different groups to be associated with bipolar disorder at the threshold of $P < 10^{-7}$ (Table 10.1).

Another GWAS study on bipolar disorder was conducted by the Psychiatric GWAS Consortium in 2011 (Ripke et al.). In this study, 7,481 European ancestry cases and 9,250 European ancestry controls were genotyped in the discovery stage of the study and then replicated in 4,496 European ancestry cases and 42,422 European ancestry controls. Eighteen significant SNPs in the initial phase of the study were replicated. Among them, two SNPs located in *CACNA1C* and *ODZ4* showed strong evidence of association ($P < 10^{-7}$). In addition, a combined analysis of schizophrenia and bipolar disorder yielded strong association evidence for SNPs in *CACNA1C* and in the region of *NEK4–ITIH1–ITIH3–ITIH4*.

Table 10.1 Genes associated with cancers or mental disorders in GWAS studies[a]

Complex disease	Associated genes/loci in GWAS studies
Cancer	
Breast cancer	*FGFR2, TOX3-CHD9, TNP1-DIRC, C10orf1, C19orf62, C6orf97-ESR1, CDKN2BAS, COL1A1, EMBP1, FGF10-MRPS30, IFITM9P-CCND1, LSP1, RNF146, RPL26P19-MAP3K1, RPL31P43-RPL36P14, SLC4A7, SRRM1P1-POU5F1B, TERT, ZMIZ1*, and *ZNF365*
Prostate cancer	*AGAP7-RPL23AP61, AR-OPHN1, C2orf43, CALM2P1-SOX9, CCHCR1, COL6A3-MLPH, CXorf67-NUDT11, DPF1-PPP1R14A, EEFSEC, EHBP1, FABP5L1-KLF12, FAM84B-SRRM1P1, FGF10, FOXP4, FSHR, GGCX-VAMP8, HNF1B, IRX4-IRX2, ITGA6, KLK3-KLK2, KRT78-RPL7P41, LMTK2, LOC100289162, MLPH, MSMB, NUDT11-TRNAE37P, PDLIM5, POU5F1B-MYC, RFX6, RPL19P16-FGFR2, RPL6P14-TET2, RPS25P10-BIK, SKIL-CLDN11, SLC22A3, SLC25A37-NKX3-1, SQRDL-SEMA6D, SRRM1P1-POU5F1B, TERT, TH-ASCL2, THADA, TPCN2-MYEOV, TUBA1C-PRPH, VGLL3-CHMP2B, ZBTB38, ZNF652*
Lung cancer	*AGPHD1, BAT3, CHRNA3, CLPTM1L, TERT, TP63, C3orf21, DYNC2H1-PDGFD, EIF4E2, MTMR3, NOP56P1-RPL13P, TNFRSF19-MIPEP*
Pancreatic cancer	*ABO, BACH1, C10orf84-PRLHR, DAB2, FABP5L1-KLF12, FABP5L1-KLF12, FAM19A5, IL17F, NR5A2, TFF2-TFF1*
Gastric cancer	*CHEK2, PLCE1, PRKAA1, ZBTB20*
Hepatocellular carcinoma	*HLA-DQB1–HLA-DQA2, HLA-S-MICA, KIF1B*
Mental disorder	
Bipolar disorder	*CACNA1C, ODZ4, ANK3-ARL4P, CACNA1C, CYCSP16-PNRC1, DGKH, HLA-B–DHFRP2, HSPD1P6-TRANK1, LMAN2L, NCAN, ODZ4, PALB2, PDE10A-C6orf176, PPM1M, RPL23AP39-RPL21P17, RPLP0P5-MARK1, TLR4-DBC1*
Schizophrenia	*BRP44, CNNM2, CSMD1, HIST1H2AH-RPL10P2, HLA-DRB1–HLA-DQA1, ITIH4, LSM1, MAP1LC3P-TCF4, NKAPL, NOTCH4, NT5C2, PRSS16-TRNAI28P, RPL26P9-FLJ35409, RPL34P22-TSPAN18, RPS17P8-GLULP6, SDCCAG8, SPA17-NRGN, TCF4, and TRIM26; 1p21.3, 2q32.3, 8p23.2, 8q21.3 and 10q24.32–q24.33, 6p21.32–p22.1 (MIR137), and 18q21.2*
Autism disorder	*MACROD2,MSNL1-CDH9*
Alzheimer's disease	*ABCA7, APOC1, BIN1-CYP27C1, CD2AP, CD33, CLU, CLU-SCARA3, CR1, EDAR-SH3RF3, EPHA1-TAS2R62P, GAB2, MS4A4E-MS4A4A, MS4A6A, MTHFD1L, PICALM-FNTAL1*, *PVRL2*, and *TOMM40*

[a]Summarized based on NHGRI GWAS Catalo, accessed by 28 Jan 2012; the catalogue includes 1,160 publications and 5,768 SNPs. URL: http://www.genome.gov/GWAStudies/index.cfm

At least 19 genes or loci have been suggested to have significant associations with schizophrenia (Table 10.1). A recent GWAS study revealed signals on 1p21.3, 2q32.3, 8p23.2, 8q21.3 and 10q24.32–q24.33, 6p21.32–p22.1 (*MIR137*), and 18q21.2 are associated with schizophrenia (Ripke et al. 2011). In addition, *MYO18B* (Purcell et al. 2009) and 2q37.2 and 2q34 (Stefansson et al. 2009) were also proposed as susceptibility loci for schizophrenia.

The susceptibility loci for other psychiatric disorders have also been identified, such as *MACROD2* (Anney et al. 2010) and the region between *MSNL1 and CDH9* gene (Wang et al. 2009) for autistic disorder, and 18 genes for Alzheimer's disease (Table 10.1).

10.3 The Implications from GWAS Findings on the *HLA* Region: Pleiotropy of Genes Among Complex Disorders

A gene is defined as pleiotropic when it influences multiple phenotypic traits, and pleiotropy is a common property of genes (Sivakumaran et al. 2011). For instance, mutations in the *PAH* gene for phenylketonuria can cause mental retardation, hair loss, pigmentation abnormalities, etc. Likewise, the tumour suppressor gene p53, mutated in many different cancers, can also be considered pleiotropic.

It is well established that multiple cancers share the same pathway of activated oncogenes, including *Ras, PI(3)K, mTOR*, *NF-kappa B, and others* (Eccleston and Dhand 2006). At least seven major signalling pathways are found in both cancer and stem cells, including *JAK/STAT, Notch, MAPK/ERK, PI3K/AKT, NF-kappa B, Wnt,* and *TGF-beta* pathways (Dreesen and Brivanlou 2007), suggesting same pathways can be involved in both disease and normal development. These key common pathways provide attractive targets for drug design.

The same concept could be applied to multiple mental disorders. Many psychiatric disorders are closely correlated to one another (also termed co-aggregated) because of their common neurobiological basis in the human brain. For instance, autism and familial major mood disorder, both of which show strong heritability, are etiologically and genetically correlated (DeLong 2004). Familial correlations and aggregations of multiple mental disorders are also observed in eating disorders and mood disorders (Mangweth et al. 2003), bipolar disorder, and schizophrenia (Van Snellenberg and de Candia 2009).

It has been proposed that there may be a link between variation of the *HLA* genes and virus infections in both cancers and psychiatric disorders. To date, at least four GWAS studies show a statistically significant association between schizophrenia and the *HLA* region, including the ISC consortium (Purcell et al. 2009), the SGENE-plus consortium (Stefansson et al. 2009), the MGS consortium (Shi et al. 2009), and a recent study by Ripke et al. (2011).

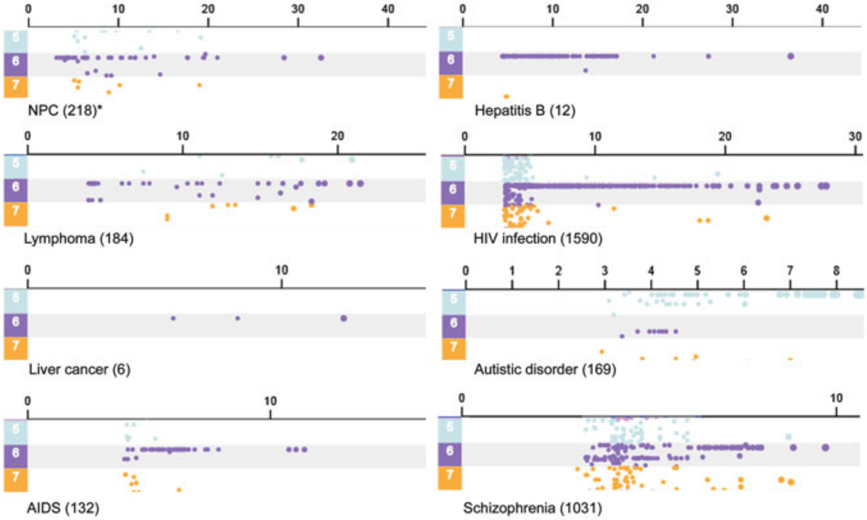

Fig. 10.1 *HLA* loci are associated with multiple human diseases related to virus infection. Note, indicates disease and number of loci with *p*-value <0.001. x-axis indicates the plot of −log10 (*p*-value) by chromosome, vertical y-axis indicates the chromosome number, only chromosome 1–chromosome 7 are shown. These plots are modified from the plots that generated in the GWASdb database (URL:http://jjwanglab.org:8080/GWASdb, Li et al. 2011b)

Likewise, GWAS studies of cancer have revealed that the *HLA* region is associated with many virus-related human cancers. Examples include Kaposi's sarcoma, thyroid carcinomas, ovarian/cervical cancer, herpes lymphoma, and liver cancer. GWAS studies with large samples of nasopharyngeal carcinoma (NPC) in populations in southern China (*HLA*-A) (Bei et al. 2010) and Taiwan (Tse et al. 2009) confirmed that *HLA* region (*HLA-A*, *HLA-F*, and *GABBR1* within chromosome 6p21.3) is significantly associated with NPC risk, confirming Simons' first finding in the early 1970s (Simons 2011; Simons and Day 1977; Simons et al. 1975, 1976, 1977). Nasopharyngeal carcinoma is characterized by Epstein–Barr virus infections and is prevalent in the Cantonese population.

In fact, *HLA* is associated with many other diseases related to viral infections (Fig. 10.1). Autistic disorder and schizophrenia, both associated with *HLA,* might also be virus related. It has been suggested that congenital rubella (MMR), hepatitis B, measles virus, and other viruses might contribute to autistic disorder (Le Blanc et al. 2003; Gallagher and Goodman 2010), and Chlamydophila psittaci, human endogenous retrovirus W, Chlamydophila pneumoniae, Borna disease virus, Toxoplasma gondii, and human herpes virus 2 might contribute to schizophrenia (Arias et al. 2011). Future studies will be needed to confirm hypotheses regarding the role of the immune system in human complex diseases.

10.4 Limitations of GWAS and the Rise of Next-Generation Sequencing Approach

Clearly, GWAS has provided a valuable tool for studying the associations between common genetic variants and human diseases. Many hundreds of SNPs have been identified that show significant associations with elevated cancer risk. Of the top 434 significant SNPs identified in 162 GWAS studies in the period of 2005–2008, 40% are located in introns of genes, 50% map to regions without any reported genes, and only 10% are within genes. The vast majority of variants identified by GWAS seem to be markers for something that are near causal factors but are not causal themselves. Criticisms have been raised about the limitations of GWAS studies, many of which have been well addressed in Chap. 4 by Jorgensen (see Sect. 4.5 in this book).

The limitations of GWAS have motivated human geneticists to reconsider the contribution of CNVs, SVs, gene–environment, gene–gene interactions, and rare variants in the pathogenesis of complex diseases. The 'common disease, common variants' hypotheses have evolved into the 'common disease, rare variants' hypotheses. Fortunately, the development of revolutionary high-throughput sequencing technologies has made this idea testable. Several large-scale sequencing projects for various cancers have been initiated, including the Cancer Genome Project (CGP; http://www.sanger.ac.uk/genetics/CGP) at the Wellcome Trust Sanger Institute in the United Kingdom, The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov) by the National Cancer Institute in the United States, and a campaign launched by the International Cancer Genome Consortium (ICGC, RUL: http://www.icgc.org) involving 27 sequencing projects in Asia, Australia, Europe, and America. These projects aim to obtain a comprehensive description of genomic, transcriptomic, and epigenomic changes in 50 different tumours. In the near future, ICGC will release a map of somatic mutations for each specific cancer in the study.

10.5 Next-Generation Sequencing Platforms and Statistical Models for Risk and Outcome Prediction

The promise of next-generation sequencing (NGS) technologies is that they can provide many orders of magnitude more data than the original Sanger sequencing methods developed in the 1970s and used to generate the first human genome sequences. In general, NGS sequencing has a higher error rate than conventional methods, but this is compensated for by greater read depth (i.e. more independent reads per position). Examples of these instruments are those made by Roche, Illumina, and Life Technologies. A similar sequencing method, available only as a service, is provided by Complete Genomics. While each company touts the advantages of their instrument and technology, they are all limited to reads of a few hundred bases, and each has particular biases in the data. A third generation of NGS technologies based on sequencing true single molecules is also being developed by companies such as Pacific Biosciences and Oxford Nanopore, but it remains to be seen if they will replace existing methods or fill specific niches such as longer reads for contig assembly.

Because of the cost benefits of NGS and the clever applications that have been developed (e.g. targeted capture using biotinylated probes), these methods have largely displaced standard sequencing and may soon be used as a substitute for genome-wide association arrays. Five major uses for NGS are (1) 'whole' genome sequencing (WGS), (2) 'whole' exome sequencing (WES), (3) transcript sequencing (i.e. RNA-Seq), (4) sequencing of genome isolated by chromatin immunoprecipitation (ChIP-Seq), and (5) methylation sequencing (bisulphite sequencing) ['whole' in quotation marks because neither method actually captures the entirety of the genome or exome]. The variety of applications continues to grow, and, collectively, these methods are giving new insights into disease mechanisms. A large number of scientific papers have been published using NGS, and this is likely to increase as costs continue to drop with the goal of the $1000 (per) genome expected shortly (Meyerson et al. 2010; Wong et al. 2011; Davey et al. 2011).

One of the main challenges of NGS in cancer studies is to determine the driving mutations among the thousands to millions of passenger mutation mutations seen in any individual cancer cell. Over 20 statistical analysis approaches have been developed for rare variant association studies, and dozens of bioinformatic tools have been developed for functional assessment of sequence variations. These issues have been well reviewed in the literature (Bansal et al. 2010). Among the statistical approaches, collapsing methods have been demonstrated to be powerful. They have also been developed to accommodate quantitative traits and applied to various designs, including family-based studies.

As genome sequencing becomes a clinical tool, reliable statistical models will be needed to link the genetic alterations to treatment strategies and disease outcomes. Building models based on machine learning methods, that is, random forests or SVM (support vector machines), might provide good prediction performances. The NGS-based statistical models should be trained on large cohorts in order to gain sufficient specificity and sensitivity for prediction. The genetic models alone or when combined with epidemiological data (e.g. smoking) and traditional clinical data (e.g. staging) should increase the performance of risk and outcome prediction.

10.6 Application of Next-Generation Sequencing to Human Complex Diseases

To date, several cancers have been investigated by NGS, including renal carcinoma, AML (acute myelogenous leukaemia), lung cancer, breast cancer, melanoma, and small-cell lung cancer (Table 10.2). Li et al. (2011a) sequenced ~18,000 genes in 10 associated hepatitis C hepatocellular carcinomas with validation in 106 tumour samples to show that mutations in the *ARID2* gene occurred in 18.2% of individuals

Table 10.2 Recent studies on cancers pursuing casual rare variants using next-generation sequencing technology

Cancer type	Sample	Method[a]	Reference
Renal carcinoma	7	ES	Varela et al. (2011)
Hepatocellular carcinoma	47	ES	Harring et al. (2011)
MSS and MSI colorectal cancer	4	ES	Timmermann et al. (2011)
Breast cancer cell line HCC1954 and a lymphoblast cell line from the same individual, HCC1954BL	2	ES and TS	Varela et al. (2010)
Metastasizing uveal melanomas	2 pairs	ES	Harbour et al. (2010)
Child fatal classic Kaposi' sarcoma	1	ES	Byun et al. (2010)
Nephronophthisis-related ciliopathies	10	ES	Otto et al. (2010)
Terminal osseous dysplasia (TOD)	6	ES	Sun et al. (2010)
AML (acute myelogenous leukaemia)	1	WGS	Ko et al. (2010)
Glioblastoma			Cancer Genome Atlas Research Network (2008)
Lung cancer	2	WGS	Campbell et al. (2008)
Breast cancer	24	WGS	Stephens et al. (2009)
Melanoma	1	WGS	Pleasance et al. (2009a)
Small-cell lung cancer	1	WGS	Pleasance et al. (2009b)
AML (acute myelogenous leukaemia)	1	WGS	Pleasance et al. (2010)
Breast cancer	1	WGS	Shah et al. (2009b)
Breast cancer	1	WGS	Ding et al. (2010)
Lung cancer	1	WGS	Lee et al. (2010)
Granulosa-cell tumours of the ovary	15	TS	Shah et al. (2009a)
Melanoma	11	TS	Berger et al. (2010)
Chronic myelogenous leukaemia cell line and prostate cancer	6	TS	Maher et al. (2009)
CCL-243 *K-562 cell line*	1	TS	Levin et al. (2009)
Oestrogen-receptor-alpha-positive metastatic lobular breast cancer	1	TS	Shah et al. (2009b)
B-cell lymphomas	2	TS	Morin et al. (2010)
Prostate cancers	25	TS	Pflueger et al. (2010)

[a] *ES* exome sequencing, *WGS* whole genome sequencing, *TS* transcriptome sequencing

with HCV-associated HCC in the United States and Europe. Agrawal et al. (2011) used the same approach to study 32 HNSCC and found that 47% patients had mutations in *TP53* and 15% patients had mutations in *NOTCH1*. Using NGS, Lee identified a wide variety of somatic variants, including more than 50,000 high-confidence single nucleotide variants in a patient with lung cancer. In particular, they found one in the known oncogene *KRAS*, 391 in other coding regions, and 43 structural variations in the whole genome (Lee et al. 2010).

NGS has also been applied to several mental disorders. Xu et al. (2011) sequenced 53 sporadic cases with schizophrenia and identified 40 *de novo* mutations in 27 cases. One of these mutations, *DGCR2*, is located in the previously defined schizophrenia-predisposing microdeletion region (22q11.2) (Xu et al. 2011). O'Roak et al. (2011) using exome sequencing identified severe *de novo* mutations in sporadic autism spectrum disorders. Using transcriptomic analysis, Voineagu et al. demonstrated the existence of dysregulated splicing of *A2BP1* (*FOX1*)-dependent alternative exons in the autistic brain. The results of their study were consistent with GWAS data and gene expression analysis (Voineagu et al. 2011).

Notably, at least five studies have used NGS to identify novel genes involved in intellectual disabilities (Topper et al. 2011). Caliskan et al. (2011) identified a novel mutation for autosomal recessive non-syndromic mental retardation in the *TECR* gene. Najmabadi et al. (2011) identified 50 novel genes for recessive cognitive disorders by sequencing 136 consanguineous families with autosomal recessive intellectual disability. Edvardson et al. (2009) sequenced 13 patients from eight Ashkenazi Jewish families and identified a homozygous mutation in the *TMEM216* gene in all affected individuals. Krawitz et al. (2010) identified *PIGV* mutations in hyperphosphatasia mental retardation syndrome. Vissers et al. identified unique non-synonymous *de novo* mutations in nine genes for mental retardation (2010).

Needless to say, NGS has opened up a new era for understanding how environmental factors alter the human genome and paved the way to a better understanding of origins of human complex diseases (Pfeifer and Hainaut 2010).

10.7 Impacts on Disease Diagnosis and Classification

One of the biggest impacts of NGS may be on the diagnosis of complex diseases. Currently, classification of cancer type and stage is largely based on pathology characterizations, that is, histological examination of tumour sections. Molecular classification based on biomarkers has shown high accuracy and has been successfully applied to several cancers. For instance, breast cancer can be classified into several subtypes by the expression of various biomarkers, including ER$^+$ (luminal A, luminal B) and ER$^-$. ER$^-$ can be further divided into two subtypes: HER2$^+$ and basal-like (ER$^-$, PR$^-$, and HER2$^-$, also called triple-negative breast cancer). Targeting cancer biomarkers can be very helpful in treatment. For example, a monoclonal antibody targeted against the HER2 biomarker – Herceptin® (trastuzumab) – is in clinical use. Also, diagnostic kits based on several key genes are already commercially available for clinical practice (i.e. Oncotype DX™, MammaPrint®, THEROS Breast Cancer Index℠, Breast Onc Px™, and MapQuant Dx™). Cancer classifications based on gene expression profiling (also called gene expression signatures) by microarray technologies or RNA-Seq shows even higher accuracy for breast cancer classification than standard biomarkers (Sorlie et al. 2001). Gene expression signatures have been proposed in various types of cancers over the last 10 years, and it was recently suggested that a classification based on miRNAs might have highly predictive and be superior to protein biomarkers.

Profiling genome-wide genetic or epigenetic alterations could lead to the discovery of more specific markers for the diagnosis of human complex diseases. In the last several years, more than a thousand loci have been shown to be significantly associated with cancers by SNParray in GWAS studies. However, only a few of them have been explained by biological studies, and most of the SNPs lack predictive value for cancer diagnosis or classification due to the small effect sizes. The limitation of currently available genetic markers means that most markers are functionally distal from the underlying mechanism However, using statistical methods, we are able to identify common genetic alterations that characterize certain types of cancers in the absence of their mechanism. This strategy is much like the protein biomarkers identified in cancers, such as the PSA level in prostate cancer and AFP in liver cancer, both of which are used as diagnosis markers without a clear understanding of their biological significance. Information about genetic or epigenetic changes could be used for complex disease classification and diagnosis similar to the traditional protein biomarkers (Esteller 2008). For instance, the hypermethylated gene *GSTP1* is potentially useful for prostate cancer diagnosis and has already undergone early clinical testing (Baden et al. 2009, 2011; Vener et al. 2008).

In contrast to association studies, NGS of the whole cancer genomes, in particular the paired cancer-normal methods, is able to identify somatic mutations in actual malignancies and is more likely to reveal the mutations that drive the cancer as well as possible environmental causes based on the types of mutations that are observed. An early study was conducted in trying to use a genetic marker identified by NGS, a 6-bp deletion in *GPNTG* gene, for diagnosis of retinitis pigmentosa*,* a rare disease with high heritability causing skeletal abnormalities (Schrader et al. 2011). The application of NGS in the field of oncology also shows encouraging results. Yokoyama et al. (2011) screened a few individual from a pedigree affected with melanoma and detected one mutation in *MITF* gene (G1075A, P.E318K, rs149617956). Then they found the E318K variant was also associated with melanoma in a large case–control sample collected from Australia and in another independent case–control sample from England (Yokoyama et al. 2011). Welch et al. suggested that NGS can even identify cytogenetically invisible oncogenes (bcr3 PML–RARA) in a patient with acute promyelocytic leukaemia (APL) (Welch et al. 2011). In the same issue of the journal, Link et al. reported a novel cancer mutation (a 3-kbp deletion in *TP53* gene) by using whole genome sequencing of a patient with therapy-related AML (tAML) (Link et al. 2011). All of these studies reflect the power of the NGS in cancer diagnosis and treatment. As NGS technologies continue to improve, comprehensively charactering genetic alterations in small amounts of tissue or even circulating tumour cells could become feasible.

10.8 Novel Drug Discovery, Personalized Medicine and Prognosis Prediction Based on Next-Generation Sequencing

The discovery of drugs that specifically target cancer cells will rely on the knowledge of specific molecular pathways that are disrupted in each cancer case. Several genetically targeted drugs have already been established by cell biology studies and traditional genetic approaches. Gleevec® (imatinib), for example, binds to an active site in certain mutated tyrosine kinases, blocking their activity and leading to cancer cell apoptosis. It is likely that people could use the same strategy to design other novel drugs. For example, the *EGFR* pathway has several targets for potential cancer treatments, including *EGFR* mutations in NSCLC (non-small-cell lung cancer), *BRAF* mutations in melanoma and thyroid cancer, and *KRAS/BRAF* mutations in colorectal cancer (Table 10.3). Other potential treatments target abnormal genes created by chromosomal translocations or gene amplifications, such as ATRA (all-trans retinoic acid) treatment for the *RARA–PML* translocation in acute promyelocytic leukaemia (APL), imatinib/dasatinib treatment for *BCR–ABL* translocations in chronic myelogenous leukaemia (CML), gefitinib/erlotinib treatment for *EML–ALK* translocations in NSCLC, trastuzumab/lapatinib treatment for *HER2* gene amplification in breast cancer and in upper GI cancer, and gefitinib/erlotinib treatment for *EGFR* mutations in lung cancer (Table 10.3). NGS can not only detect previously identified mutations associated with a given cancer but can also identify novel mutations that are potential targets for further drug discovery.

Patients respond to treatment in distinct ways. Some patients might be more responsive to radiotherapy, while others might respond better to cytotoxic T-cell biotherapy. The reasons for their different responses may lie in the somatic mutations in their cancer cells or the inherited variation in their normal genomes. Personalized medicine offers the expectation that physicians would be able to make the best choice from a panel of available drugs for a specific patient or, conversely, to identify which subpopulation of patients is most suitable for a particular drug or treatment method. A small number of protein markers have been routinely used to predict drug response. For example, ER (oestrogen receptor) levels have been used as a predictor for oestrogen antagonists and c-ErBB2 for trastuzumab/trastuzumab emtansine (T-DM1) in breast cancer treatment. Recent clinical trial conducted by Genentech/Roche showed that treatment with T-DM1 resulted in a marked improvement for women with HER2-positive metastatic breast cancer. It should be possible to make better predictions based on sequence information, and we would expect as more data accumulates, there will be a transition to DNA- or RNA-based predictors (Andersen et al. 2010; Ji et al. 2009). In another example, at least 64 clinical publications have been conducted to investigate the relationship of p53 to clinical outcomes following chemotherapy in ovarian cancer cases (Hall et al. 2004). While there has been much progress in this area, clinically useful predictions based on genomic data remain challenging.

Table 10.3 FDA-approved genetically targeted drugs for cancer treatments[a]

Drug	Targeted genetic alteration	Tumour	Status
ATRA/PDQ®	RARA–PML translocation/t (15;17)(q24;q21) translocation	Acute promyelocytic leukaemia (APL)	Approved
Imatinib mesylate (Gleevec®)	BCR–ABL translocation	Chronic myelogenous leukaemia (CML)	Approved in 2001
	C-kit mutation	Gastrointestinal stromal tumours (GIST)	Approved in 2002
Dasatinib (Sprycel®)	BCR–ABL translocation	Philadelphia chromosome-positive CML	Approved in 2010
Trastuzumab (Herceptin®)	HER2 gene amplification	Breast cancer	Approved in 2006
		Metastatic gastric or gastroesophageal (GE) junction adenocarcinoma	Approved in 2010
Lapatinib ditosylate (Tykerb®)	HER2 gene amplification	Breast cancer	Approved in 2007
Panitumumab (Vectibix®)	KRAS mutation/BRAF mutation	Colorectal cancer (CRC)	Approved in 2006
Cetuximab (Erbitux®)	KRAS mutation/BRAF mutation	CRC	Approved in 2004
		Head and neck cancer(SCCHN)	Approved in 2006
Sunitinib malate (Sutent®)	C-kit mutation	GIST	Approved in 2006
		Advanced kidney cancer	Approved in 2006
Vemurafenib (Zelboraf™)	BRAF V600E mutation	Unresectable or metastatic melanoma	Approved in 2011
Erlotinib hydrochloride (Tarceva®)	EGFR mutation	Non-small-cell lung cancer (NSCLC)	Approved in 2004
		Pancreatic cancer	Approved in 2005
Gefitinib (Iressa®)	EGFR mutation	NSCLC	Approved in 2003
Crizotinib (Xalkori®)	EML–ALK translocation	NSCLC	Approved in 2011

[a] Source: National Cancer Institute, URL: http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted

The National Institute of General Medical Science (NIGMS) of the NIH has funded the Pharmacogenmics Research Network, a study to identify significant sequence variants that affect drug responses. As part of the program, a free online database, the Pharmacogenomics Knowledge Base (http://www.pharmgkb.org) has been created that lists genetic variants associated with diseases and drug responses from clinical studies. It is expected that as phargkb.org continues to catalogue variants associated with drug responses, this database will become an increasingly useful tool for physicians and their patients.

10.9 Public Health Significance

Preventive disease interventions are clearly more efficient than treating diseases after they appear. Examples include smoking control, which has dramatically reduced lung cancer prevalence; early screening for breast and colon cancer; and vaccine programs for cancer-causing viruses such as the HPV for cervical cancer.

NGS is likely to identify additional targets for intervention. More than 400 genes have been reported to be associated with intellectual disability (ID) and related cognitive disorders (CDs). Several common pathways such as synaptic plasticity, Ras and Rho GTPase signalling, and chromatin remodelling are implicated. These may provide an opportunities for future therapeutic interventions (van Bokhoven 2011). Talkowski ME et al. used exome sequencing to study a region of 2p23.1, where multiple lines of evidence demonstrated the existence of microdeletions or translocations. Subsequently, the authors identified a mutation in the methyl-CpG-binding domain in the *MBD5* gene from this region as a significant risk factor for autism spectrum disorder (ASD) and warrants consideration as a clinical marker in ASD (Talkowski et al. 2011).

10.10 Summary

We continue to discover more about how genetic variations in normal and cancer cells affect the phenotypes, though much remains to be learned. With continuing improvements in the technology, it is clear that sequencing will become an increasingly important tool in predicting the risk of disease and in discovering new approaches of prevention. If technologies such as nanopore sequencing become a reality, it is quite possible that the reduction of cost and the increase in speed of genome sequencing will move genomic analysis from the research laboratory to the clinical laboratory faster than expected. However, we must develop a new framework for coordinating basic discovery with patient medical records in a way that we can transition to the medicine of the future. Statisticians, bioinformaticists, biologists, and physicians must work together more closely than ever before to achieve this promise.

References

Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, Zhang N, El-Naggar AK, Jasser SA, Weinstein JN, Trevino L, Drummond JA, Muzny DM, Wu Y, Wood LD, Hruban RH, Westra WH, Koch WM, Califano JA, Gibbs RA, Sidransky D, Vogelstein B, Velculescu VE, Papadopoulos N, Wheeler DA, Kinzler KW, Myers JN. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;333(6046):1154–7. doi:science.1206923.

The American Cancer Society. Global cancer facts & figures 2007. Atlanta: The American Cancer Society; 2007.

American Psychiatric Association (APA). Diagnostic and statistical manual of mental disorders (DSM-IV, 4th ed., text rev.). Washington, DC: American Psychiatric Association; 2000.

Andersen JB, Factor VM, Marquardt JU, Raggi C, Lee YH, Seo D, Conner EA, Thorgeirsson SS. An integrated genomic and epigenomic approach predicts therapeutic response to zebularine in human liver cancer. Sci Transl Med. 2010;2(54):54ra77. doi:2/54/54ra77.

Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT, Almeida J, Bacchelli E, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Carson AR, Casallo G, Casey J, Chu SH, Cochrane L, Corsello C, Crawford EL, Crossett A, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Melhem NM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Piven J, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Van Engeland H, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Wing K, Wittemeyer K, Wood S, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Betancur C, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Gallagher L, Geschwind DH, Gill M, Haines JL, Miller J, Monaco AP, Nurnberger Jr JI, Paterson AD, Pericak-Vance MA, Schellenberg GD, Scherer SW, Sutcliffe JS, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Devlin B, Ennis S, Hallmayer J. A genome-wide scan for common alleles affecting risk for autism. Hum Mol Genet. 2010;19(20):4072–82. doi:ddq307.

Arias I, Sorlozano A, Villegas E, Luna JD, McKenney K, Cervilla J, Gutierrez B, Gutierrez J. Infectious agents associated with schizophrenia: a meta-analysis. Schizophr Res. 2011. doi:S0920-9964(11)00561-5.

Baden J, Green G, Painter J, Curtin K, Markiewicz J, Jones J, Astacio T, Canning S, Quijano J, Guinto W, Leibovich BC, Nelson JB, Vargo J, Wang Y, Wuxiong C. Multicenter evaluation of an investigational prostate cancer methylation assay. J Urol. 2009;182(3):1186–93. doi:S0022-5347(09)01140-9.

Baden J, Adams S, Astacio T, Jones J, Markiewicz J, Painter J, Trust C, Wang Y, Green G. Predicting prostate biopsy result in men with prostate specific antigen 2.0 to 10.0 ng/ml using an investigational prostate cancer methylation assay. J Urol. 2011;186(5):2101–6. doi:S0022-5347(11)04327-8.

Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85. doi:nrg2867.

Bartley M, Bokde AL, O’Neill D. Mild cognitive impairment. N Engl J Med. 2011;365(14):1357–8; author reply 1358–9. doi:10.1056/NEJMc1108238#SA1.

Bei JX, Li Y, Jia WH, Feng BJ, Zhou G, Chen LZ, Feng QS, Low HQ, Zhang H, He F, Tai ES, Kang T, Liu ET, Liu J, Zeng YX. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet. 2010;42(7):599–603. doi:ng.601.

Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, Garraway LA. Integrative analysis of the melanoma transcriptome. Genome Res. 2010;20(4):413–27. doi:gr.103697.109.

Byun M, Abhyankar A, Lelarge V, Plancoulaine S, Palanduz A, Telhan L, Boisson B, Picard C, Dewell S, Zhao C, Jouanguy E, Feske S, Abel L, Casanova JL. Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. J Exp Med. 2010;207(11):2307–12. doi:jem.20101597.

Caliskan M, Chong JX, Uricchio L, Anderson R, Chen P, Sougnez C, Garimella K, Gabriel SB, dePristo MA, Shakir K, Matern D, Das S, Waggoner D, Nicolae DL, Ober C. Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum Mol Genet. 2011;20(7):1285–9. doi:ddq569.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722–9. doi:ng.128.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8.

Chukwujama O, Gormley N. Mild cognitive impairment. N Engl J Med. 2011;365(14):1358; author reply 1358–9. doi:10.1056/NEJMc1108238#SA2.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12(7):499–510. doi:nrg3012.

DeLong R. Autism and familial major mood disorder: are they related? J Neuropsychiatry Clin Neurosci. 2004;16(2):199–213. doi:10.1176/appi.neuropsych.16.2.199.

Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson Jr DM, Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010;464(7291):999–1005. doi:nature08989.

Dreesen O, Brivanlou AH. Signaling pathways in cancer and embryonic stem cells. Stem Cell Rev. 2007;3(1):7–17. doi:SCR:3:1:7.

Eccleston A, Dhand R. Signalling in cancer. Nature. 2006;441(7092):423–57. doi:10.1038/441423a.

Edvardson S, Shaag A, Zenvirt S, Erlich Y, Hannon GJ, Shanske AL, Gomori JM, Ekstein J, Elpeleg O. Joubert syndrome 2 (JBTS2) in Ashkenazi Jews is associated with a TMEM216 mutation. Am J Hum Genet. 2009;86(1):93–7. doi:S0002-9297(09)00568-0.

Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358(11):1148–59. doi:358/11/1148.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010;127(12):2893–917. doi:10.1002/ijc.25516.

Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, Smoller JW, Grozeva D, Stone J, Nikolov I, Chambert K, Hamshere ML, Nimgaonkar VL, Moskvina V, Thase ME, Caesar S, Sachs GS, Franklin J, Gordon-Smith K, Ardlie KG, Gabriel SB, Fraser C, Blumenstiel B, Defelice M, Breen G, Gill M, Morris DW, Elkin A, Muir WJ, McGhee KA, Williamson R, MacIntyre DJ, MacLean AW, St CD, Robinson M, Van Beck M, Pereira AC, Kandaswamy R, McQuillin A, Collier DA, Bass NJ, Young AH, Lawrence J, Ferrier IN, Anjorin A, Farmer A, Curtis D, Scolnick EM, McGuffin P, Daly MJ, Corvin AP, Holmans PA, Blackwood DH, Gurling HM, Owen MJ, Purcell SM, Sklar P, Craddock N. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 2008;40(9):1056–8. doi:ng.209.

Fombonne E. The epidemiology of autism: a review. Psychol Med. 1999;29(4):769–86.

Gallagher CM, Goodman MS. Hepatitis B vaccination of male neonates and autism diagnosis, NHIS 1997–2002. J Toxicol Environ Health A. 2010;73(24):1665–77. doi:929161456.

Hall J, Paul J, Brown R. Critical evaluation of p53 as a prognostic marker in ovarian cancer. Expert Rev Mol Med. 2004;6(12):1–20. doi:10.1017/S1462399404007781.

Harbour JW, Onken MD, Roberson ED, Duan S, Cao L, Worley LA, Council ML, Matatall KA, Helms C, Bowcock AM. Frequent mutation of BAP1 in metastasizing uveal melanomas. Science. 2010;330(6009):1410–13. doi:science.1194472.

Harring TR, Guiteau JJ, Nguyen NT, Cotton RT, Gingras MC, Wheeler DA, O'Mahony CA, Gibbs RA, Brunicardi FC, Goss JA. Building a comprehensive genomic program for hepatocellular carcinoma. World J Surg. 2011;35(8):1746–50. doi:10.1007/s00268-010-0954-x.

Ji J, Shi J, Budhu A, Yu Z, Forgues M, Roessler S, Ambs S, Chen Y, Meltzer PS, Croce CM, Qin LX, Man K, Lo CM, Lee J, Ng IO, Fan J, Tang ZY, Sun HC, Wang XW. MicroRNA expression, survival, and response to interferon in liver cancer. N Engl J Med. 2009;361(15):1437–47. doi:361/15/1437.

Kessler RC, Wang PS. The descriptive epidemiology of commonly occurring mental disorders in the United States. Annu Rev Public Health. 2008;29:115–29. doi:10.1146/annurev.publhealth.29.020907.090847.

Kessler RC, Demler O, Frank RG, Olfson M, Pincus HA, Walters EE, Wang P, Wells KB, Zaslavsky AM. Prevalence and treatment of mental disorders, 1990 to 2003. N Engl J Med. 2005;352(24):2515–23. doi:352/24/2515.

Ko M, Huang Y, Jankowska AM, Pape UJ, Tahiliani M, Bandukwala HS, An J, Lamperti ED, Koh KP, Ganetzky R, Liu XS, Aravind L, Agarwal S, Maciejewski JP, Rao A. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature. 2010;468(7325):839–43. doi:nature09586.

Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Kohler S, Jager M, Grunhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet. 2010;42(10):827–9. doi:ng.653.

Le Blanc K, Tammik C, Rosendahl K, Zetterberg E, Ringden O. HLA expression and immunologic properties of differentiated and undifferentiated mesenchymal stem cells. Exp Hematol. 2003;31(10):890–6. doi:S0301472X03001103.

Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, Antipova A, Lee C, McKernan K, De La Vega FM, Kinzler KW, Vogelstein B, Diaz Jr LA, Velculescu VE. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. 2010;2(20):20ra14. doi:2/20/20ra14.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010;465(7297):473–7. doi:nature09004.

Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol. 2009;10(10):R115. doi:gb-2009-10-10-r115.

Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, Pawlik TM, Daniel HD, Kannangai R, Offerhaus GJ, Velculescu VE, Wang L, Zhou S, Vogelstein B, Hruban RH, Papadopoulos N, Cai J, Torbenson MS, Kinzler KW. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet. 2011a;43(9):828–9. doi:ng.903.

Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, Wong MP, Sham PC, Chanock SJ, Wang J. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. 2011b. doi:gkr1182.

Link DC, Schuettpelz LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M, Koboldt DC, Dooling DJ, Fulton RS, Bender RH, Fulton LL, Delehaunty KD, Fronick CC, Appelbaum EL, Schmidt H, Abbott R, O'Laughlin M, Chen K, McLellan MD, Varghese N, Nagarajan R, Heath S, Graubert TA, Ding L, Ley TJ, Zambetti GP, Wilson RK, Mardis ER. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA. 2011;305(15):1568–76. doi:305/15/1568.

MacKenzie TD, Kolpak SJ, Mehler PS (2005) Prevalence and treatment of mental disorders. N Engl J Med. 2005;353(11):1184; author reply 1184. doi:353/11/1184.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009;458(7234):97–101. doi:nature07638.

Mangweth B, Hudson JI, Pope HG, Hausmann A, De Col C, Laird NM, Beibl W, Tsuang MT. Family study of the aggregation of eating disorders and mood disorders. Psychol Med. 2003;33(7):1319–23.

Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010;11(10):685–96. doi:nrg2841.

Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, Paul JE, Boyle M, Woolcock BW, Kuchenbauer F, Yap D, Humphries RK, Griffith OL, Shah S, Zhu H, Kimbara M, Shashkin P, Charlot JF, Tcherpakov M, Corbett R, Tam A, Varhol R, Smailus D, Moksa M, Zhao Y, Delaney A, Qian H, Birol I, Schein J, Moore R, Holt R, Horsman DE, Connors JM, Jones S, Aparicio S, Hirst M, Gascoyne RD, Marra MA. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat Genet. 2010;42(2):181–5. doi:ng.518.

Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, Hosseini M, Behjati F, Haas S, Jamali P, Zecha A, Mohseni M, Puttmann L, Vahid LN, Jensen C, Moheb LA, Bienek M, Larti F, Mueller I, Weissmann R, Darvish H, Wrogemann K, Hadavi V, Lipkowitz B, Esmaeeli-Nieh S, Wieczorek D, Kariminejad R, Firouzabadi SG, Cohen M, Fattahi Z, Rost I, Mojahedi F, Hertzberg C, Dehghan A, Rajab A, Banavandi MJ, Hoffer J, Falah M, Musante L, Kalscheuer V, Ullmann R, Kuss AW, Tzschach A, Kahrizi K, Ropers HH. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature. 2011;478(7367):57–63. doi:nature10423.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246–50. doi:10.1038/nature10989.

Otto EA, Hurd TW, Airik R, Chaki M, Zhou W, Stoetzel C, Patil SB, Levy S, Ghosh AK, Murga-Zamalloa CA, van Reeuwijk J, Letteboer SJ, Sang L, Giles RH, Liu Q, Coene KL, Estrada-Cuzcano A, Collin RW, McLaughlin HM, Held S, Kasanuki JM, Ramaswami G, Conte J, Lopez I, Washburn J, Macdonald J, Hu J, Yamashita Y, Maher ER, Guay-Woodford LM, Neumann HP, Obermuller N, Koenekoop RK, Bergmann C, Bei X, Lewis RA, Katsanis N, Lopes V, Williams DS, Lyons RH, Dang CV, Brito DA, Dias MB, Zhang X, Cavalcoli JD, Nurnberg G, Nurnberg P, Pierce EA, Jackson PK, Antignac C, Saunier S, Roepman R, Dollfus H, Khanna H, Hildebrandt F. Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. Nat Genet. 2010;42(10):840–50. doi:ng.662.

Petersen RC. Clinical practice. Mild cognitive impairment. N Engl J Med. 2011;364(23):2227–34. doi:10.1056/NEJMcp0910237.

Pfeifer GP, Hainaut P. Next-generation sequencing: emerging lessons on the origins of human cancer. Curr Opin Oncol. 2010;23(1):62–8. doi:10.1097/CCO.0b013e3283414d00.

Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 2010;21(1):56–67. doi:gr.110684.110.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, Stratton MR. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2009a;463(7278):191–6. doi:nature08658.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordonez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2009b;463(7278):184–90. doi:nature08629.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordonez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010;463(7278):184–90. doi:nature08629.

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–52. doi:nature08185.

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DH, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jonsson EG, Bitter I, Pietilainen OP, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Borglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Haan L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthoj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jurgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krasucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang KY, Lichtenstein P, Lieberman JA, Linszen DH, Lonnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattingsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentoft M, Norton N, Nothen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Orntoft TF, Owen MJ, Pantelis C, Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Rethelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CC, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Terenius L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43(10):969–76. doi:ng.940.

Schrader KA, Heravi-Moussavi A, Waters PJ, Senz J, Whelan J, Ha G, Eydoux P, Nielsen T, Gallagher B, Oloumi A, Boyd N, Fernandez BA, Young TL, Jones SJ, Hirst M, Shah SP, Marra MA, Green J, Huntsman DG. Using next-generation sequencing for the diagnosis of rare disorders: a family with retinitis pigmentosa and skeletal abnormalities. J Pathol. 2011;225(1):12–8. doi:10.1002/path.2941.

Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, Wiegand KC, Leung G, Zayed A, Mehl E, Kalloger SE, Sun M, Giuliany R, Yorida E, Jones S, Varhol R, Swenerton KD, Miller D, Clement PB, Crane C, Madore J, Provencher D, Leung P, DeFazio A, Khattra J, Turashvili G, Zhao Y, Zeng T, Glover JN, Vanderhyden B, Zhao C, Parkinson CA, Jimenez-Linan M, Bowtell DD, Mes-Masson AM, Brenton JD, Aparicio SA, Boyd N, Hirst M, Gilks CB, Marra M, Huntsman DG. Mutation of FOXL2 in granulosa-cell tumors of the ovary. N Engl J Med. 2009a;360(26):2719–29. doi:NEJMoa0902542.

Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009b;461(7265):809–13. doi:nature08489.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009;460(7256):753–7. doi:nature08192.

Simons MJ. Nasopharyngeal carcinoma as a paradigm of cancer genetics. Chin J Cancer. 2011;30(2):79–84. doi:1944-446X20110279.

Simons MJ, Day NE. Histocompatibility leukocyte antigen patterns and nasopharyngeal carcinoma. Natl Cancer Inst Monogr. 1977;47:143–6.

Simons MJ, Wee GB, Chan SH, Shanmugaratnam K, Day NE, de-The G. Immunogenetic aspects of nasopharyngeal carcinoma (NPC) III. HL-a type as a genetic marker of NPC predisposition to test the hypothesis that Epstein-Barr virus is an etiological factor in NPC. IARC Sci Publ. 1975;11(Pt 2):249–58.

Simons MJ, Wee GB, Goh EH, Chan SH, Shanmugaratnam K, Day NE, de-The G. Immunogenetic aspects of nasopharyngeal carcinoma. IV. Increased risk in Chinese of nasopharyngeal carcinoma associated with a Chinese-related HLA profile (A2, Singapore 2). J Natl Cancer Inst. 1976;57(5):977–80.

Simons MJ, Wee GB, Singh D, Dharmalingham S, Yong NK, Chau JC, Ho JH, Day NE, De-The G. Immunogenetic aspects of nasopharyngeal carcinoma. V. Confirmation of a Chinese-related HLA profile (A2, Singapore 2) associated with an increased risk in Chinese for nasopharyngeal carcinoma. Natl Cancer Inst Monogr. 1977;47:147–51.

Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet. 2011;89(5):607–18. doi:S0002-9297(11)00438-1.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869–74. doi:10.1073/pnas.191367098.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Borglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Bottcher Y, Olesen J, Breuer R, Moller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Rethelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jonsson EG, Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA. Common variants conferring risk of schizophrenia. Nature. 2009;460(7256):744–7. doi:nature08186.

Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW, Burton J, Quail MA, Swerdlow H, Churcher C, Natrajan R, Sieuwerts AM, Martens JW, Silver DP, Langerod A, Russnes HE, Foekens JA, Reis-Filho JS, Richardson AL, Borresen-Dale AL, Campbell PJ, Futreal PA, Stratton MR, Van't Veer L. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009;462(7276):1005–10. doi:nature08645.

Sun Y, Almomani R, Aten E, Celli J, van der Heijden J, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L, Horii E, Drut R, Ariyurek Y, den Dunnen JT, Breuning MH. Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet. 2010;87(1):146–53. doi:S0002-9297(10)00311-3.

Talkowski ME, Mullegama SV, Rosenfeld JA, van Bon BW, Shen Y, Repnikova EA, Gastier-Foster J, Thrush DL, Kathiresan S, Ruderfer DM, Chiang C, Hanscom C, Ernst C, Lindgren AM, Morton CC, An Y, Astbury C, Brueton LA, Lichtenbelt KD, Ades LC, Fichera M, Romano C, Innis JW, Williams CA, Bartholomew D, Van Allen MI, Parikh A, Zhang L, Wu BL, Pyatt RE, Schwartz S, Shaffer LG, de Vries BB, Gusella JF, Elsea SH. Assessment of 2q23.1 Microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. Am J Hum Genet. 2011;89(4):551–63. doi:S0002-9297(11)00401-0.

Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, Wunderlich A, Barmeyer C, Seemann P, Koenig J, Lappe M, Kuss AW, Garshasbi M, Bertram L, Trappe K, Werber M, Herrmann BG, Zatloukal K, Lehrach H, Schweiger MR. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One. 2011;5(12):e15661. doi:10.1371/journal.pone.0015661.

Topper S, Ober C, Das S. Exome sequencing and the genetics of intellectual disability. Clin Genet. 2011;80(2):117–26. doi:10.1111/j.1399-0004.2011.01720.x.

Tse KP, Su WH, Chang KP, Tsang NM, Yu CJ, Tang P, See LC, Hsueh C, Yang ML, Hao SP, Li HY, Wang MH, Liao LP, Chen LC, Lin SR, Jorgensen TJ, Chang YS, Shugart YY. Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am J Hum Genet. 2009;85(2):194–203. doi:S0002-9297(09)00298-5.

van Bokhoven H. Genetic and epigenetic networks in intellectual disabilities. Annu Rev Genet. 2011;45:81–104. doi:10.1146/annurev-genet-110410-132512.

Van Snellenberg JX, de Candia T. Meta-analytic evidence for familial coaggregation of schizophrenia and bipolar disorder. Arch Gen Psychiatry. 2009;66(7):748–55. doi:66/7/748.

Varela I, Klijn C, Stephens PJ, Mudie LJ, Stebbings L, Galappaththige D, van der Gulden H, Schut E, Klarenbeek S, Campbell PJ, Wessels LF, Stratton MR, Jonkers J, Futreal PA, Adams DJ. Somatic structural rearrangements in genetically engineered mouse mammary tumors. Genome Biol. 2010;11(10):R100. doi:gb-2010-11-10-r100.

Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D, Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LF, Richard S, Kahnoski RJ, Anema J, Tuveson DA, Perez-Mancera PA, Mustonen V, Fischer A, Adams DJ, Rust A, Chan-on W, Subimerb C, Dykema K, Furge K, Campbell PJ, Teh BT, Stratton MR, Futreal PA. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011;469(7331):539–42. doi:nature09639.

Vener T, Derecho C, Baden J, Wang H, Rajpurohit Y, Skelton J, Mehrotra J, Varde S, Chowdary D, Stallings W, Leibovich B, Robin H, Pelzer A, Schafer G, Auprich M, Mannweiler S, Amersdorfer P, Mazumder A. Development of a multiplexed urine assay for prostate cancer diagnosis. Clin Chem. 2008;54(5):874–82. doi:clinchem.2007.094912.

Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. A de novo paradigm for mental retardation. Nat Genet. 2010;42(12):1109–12. doi:ng.712.

Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380–4. doi:nature10110.

Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM, Kim CE, Hou C, Frackelton E, Chiavacci R, Takahashi N, Sakurai T, Rappaport E, Lajonchere CM, Munson J, Estes A, Korvatska O, Piven J, Sonnenblick LI, Alvarez Retuerto AI, Herman EI, Dong H, Hutman T, Sigman M, Ozonoff S, Klin A, Owley T, Sweeney JA, Brune CW, Cantor RM, Bernier R, Gilbert JR, Cuccaro ML, McMahon WM, Miller J, State MW, Wassink TH, Coon H, Levy SE, Schultz RT, Nurnberger JI, Haines JL, Sutcliffe JS, Cook EH, Minshew NJ, Buxbaum JD, Dawson G, Grant SF, Geschwind DH, Pericak-Vance MA, Schellenberg GD, Hakonarson H. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009;459(7246):528–33. doi:nature07999.

Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, Wallis J, Chen K, Payton JE, Fulton RS, Veizer J, Schmidt H, Vickery TL, Heath S, Watson MA, Tomasson MH, Link DC, Graubert TA, DiPersio JF, Mardis ER, Ley TJ, Wilson RK. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA. 2011;305(15):1577–84. doi:305/15/1577.

Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. Annu Rev Genomics Hum Genet. 2011;12:407–30. doi:10.1146/annurev-genom-082509-141532.

WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78. doi:nature05911.

Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M. Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet. 2011;43(9):864–8. doi:ng.902.

Yokoyama S, Woods SL, Boyle GM, Aoude LG, MacGregor S, Zismann V, Gartside M, Cust AE, Haq R, Harland M, Taylor JC, Duffy DL, Holohan K, Dutton-Regester K, Palmer JM, Bonazzi V, Stark MS, Symmons J, Law MH, Schmidt C, Lanagan C, O'Connor L, Holland EA, Schmid H, Maskiell JA, Jetann J, Ferguson M, Jenkins MA, Kefford RF, Giles GG, Armstrong BK, Aitken JF, Hopper JL, Whiteman DC, Pharoah PD, Easton DF, Dunning AM, Newton-Bishop JA, Montgomery GW, Martin NG, Mann GJ, Bishop DT, Tsao H, Trent JM, Fisher DE, Hayward NK, Brown KM. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. Nature. 2011;480(7375):99–103. doi:10.1038/nature10630.

Index

Y.Y. Shugart (ed.), *Applied Computational Genomics*, Translational Bioinformatics 1, DOI 10.1007/978-94-007-5558-1, © Springer Science+Business Media Dordrecht 2012