Translational Bioinformatics 13
Series Editor: Xiangdong Wang, MD, PhD, Prof

Yin Yao *Editor*

Applied Computational Genomics

Second Edition

Springer

Translational Bioinformatics

Volume 13

Series editor
Xiangdong Wang, MD, Ph.D.
Professor of Medicine, Zhongshan Hospital, Fudan University Medical School, China
Director of Shanghai Institute of Clinical Bioinformatics, (www.fuccb.org)

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.

Recently Published and Forthcoming Volumes

Allergy Bioinformatics
Editors: Ailin Tao, Eyal Raz
Volume 8

Transcriptomics and Gene Regulation
Editor: Jiaqian Wu
Volume 9

Pediatric Biomedical Informatics - Computer Applications in Pediatric Research (Edition 2)
Editor: John J. Hutton
Volume 10

Application of Clinical Bioinformatics
Editors: Xiangdong Wang, Christian Baumgartner, Denis C. Shields, Hong-Wen Deng, Jacques S Beckmann
Volume 11

More information about this series at http://www.springer.com/series/11057

Yin Yao
Editor

Applied Computational Genomics

Second Edition

Springer

Editor
Yin Yao
Intramural Research Program,
National Institute of Mental Health
National Institutes of Health
Bethesda, MD, USA

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-981-13-1070-6 ISBN 978-981-13-1071-3 (eBook)
https://doi.org/10.1007/978-981-13-1071-3

Library of Congress Control Number: 2018950448

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Contributors

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

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

KeyBank, Cleveland, OH, USA

Li Jin Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

Yi Li Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

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

Haide Qin Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, People's Republic of China

McKenzie Ritter Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA

Yi Wang Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

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

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

Chapter 1
Introduction

McKenzie Ritter, Yin Yao, and Andrew Collins

Abstract This chapter presents an overview of the current genomic field and highlights each of the ten chapters, which have been collected into this book. The critical concepts illustrated by the authors are also pointed out through logical connections between different chapters.

1.1 Overview

With the recent use of both large and multidimensional data, a demand has been created for the development of new tools that can properly analyze it. Applied statistical genomics methodologies have been developed to handle just that. This book presents various analyses that deal with such complex data, which contrasts from previous analyses used in the field, such as linkage and segregation analyses. With the more common use of large data, analyses such as sequencing are extremely important to the field. The goal of this book is to present these various examples of analyses in order to provide other researchers with the tools to utilize and apply them to their own genomic data.

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 improves the efficiency and infrastructure for data management, and (3) the

M. Ritter · Y. Yao
Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA

A. Collins (✉)
Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, University of Southampton, Southampton, UK
e-mail: a.r.collins@soton.ac.uk

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_1

development of more powerful statistical tools to analyze 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 investigators with fresh knowledge to 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 in the context of techniques, such as genome-wide association studies (GWAS), family-based linkage analyses, candidate gene-based approaches, and case-control-based association analyses. 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 etiology 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 provide a clear route toward greatly improved clinical application of these new sources of data. 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 Overview of Chapter Contents

This book has 11 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: this chapter serves as an introduction to the current status of the genomic field and provides highlights of the ten 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 opens a door to the important statistical tools used by 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, the methodology and results of a whole genome analysis on the overlap between attention deficit hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) are discussed. A GWAS, meta-analysis, polygenic risk score analysis, protein-protein link evaluation, and expression quantitative trait locus (eQTL) analysis are conducted on the two data sets to search for overlapping genetic variants that may point to the susceptibility of the two disorders. Each of these analyses is discussed, as well as the results from each of the analyses. This chapter is based on a previously published manuscript (Ritter et al. 2017).

In Chap. 3, a complete review of the most important concepts in genetic epidemiology is provided. This chapter moves beyond the traditional risk factors defined by epidemiologists and reviews breakthroughs in genomics in recent years. A precise definition for complex traits is provided, as well as 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, and migration are also reviewed, as well as 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. This work provides a strong foundation for the remainder of the book.

The in-depth statistical framework for rare variants analysis can be found in Chap. 4, in which all currently analytical strategies on rare variant hunting are discussed. The idea of "collapsing" is explained and then provides 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 ratio weighted sum statistic (ORWSS) (Price et al. 2010). Furthermore, 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 is provided.

Chapter 5 discusses a whole genome association analysis on the treatment response of individuals with OCD. This chapter is based on an analysis that was conducted and published previously (Qin et al. 2015). Up to 30% of individuals suffering from OCD are treatment resistant to serotonin reuptake inhibitors (SRIs). Thus, this analysis sought to identify genetic predictors of treatment response. The enrichment analysis indicated that two pathways were significant: the glutamatergic neurotransmission pathway and serotonergic neurotransmission pathway. This was the first GWAS to examine treatment response of OCD.

Chapter 6 serves as 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 analyzed 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. It is concluded that although eQTL mapping in the human brain is in its early stages, as a tool, QTL has the potential to identify important disease intermediate phenotypes, as well as a route to further understanding complex diseases. Furthermore, the author describes several commonly used experimental platforms and analytical procedures related to eQTL

studies. All literature is also reviewed on mQTL, in which DNA methylation levels at specific CpG sites are considered as quantitative traits. More importantly, a list of databases is included for QTL mapping results that were built by scientists who collectively have collated basic scientific knowledge, enabling the advancement of personal medicine.

Chapter 7 examines research progress in a specific rare disorder. This disease is nasopharyngeal carcinoma (NPC). A thorough review of all candidate genes related to NPC was conducted (note, this was limited to work published in English), as well as commentary 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 1500 cases and 1500 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 a 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 suspect 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.

Chapter 8 discusses a method that tests for nonlinear dependence of two continuous variables. This was termed CANOVA and is named so because of its basis from ANOVA but differs in its use of continuous variables. This method was previously developed and published (Wang et al. 2012). The proposed method was compared to six other existing methods through the use of simulation data and RNA-seq kidney cancer data. The CANOVA method held up against the other six methods it was compared to but does have some limitations. It would be a good idea to test this method further using other real data.

Chapter 9 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. It is also stated 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, it is noted 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 main statistical methods that can be used to analyze both

rare variants and CNVs are introduced. Readers who are eager to grasp analytical concepts relating to de novo variants, the behaviors 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.

Chapter 10 provides a brief introduction to machine learning. Additionally, the applications of machine learning in the context of genomic data are discussed. Machine learning can be used for both disease prediction and disease prognosis. The limits of machine learning as well as the hopes of its applications for the future are mentioned.

The last chapter, Chap. 11, covers a very important area of human genetics, which is the history and progression of genetic analyses. The progression of the field is discussed, beginning with segregation analyses and then discussing linkage analyses, GWAS, and meta-analyses. Each of the analyses is explained and then placed in the context of OCD with results discussed from published studies in each of the mentioned analyses. The future of OCD genomics is also discussed, which includes the use of rare variants, large sample sizes, and more precise phenotype classification. This chapter wraps up by summarizing the most common analyses in the field and also looks to the future and suggests several focuses to increase the power and clarity of the studies.

References

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

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

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.

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.

Qin H, Samuels JF, Wang Y, Zhu Y, Grados MA, Riddle MA, et al. Whole-genome association analysis of treatment response in obsessive-compulsive disorder. Mol Psychiatry. 2015;21:270–6. https://doi.org/10.1038/mp.2015.32. Macmillan Publishers Limited.

Ritter ML, Guo W, Samuels JF, 2012 Y, Nestadt PS, Krasnow J, et al. Genome wide association study (GWAS) between attention deficit hyperactivity disorder (ADHD) and obsessive compulsive disorder (OCD). Front Mol Neurosci. 2017;10(83). https://doi.org/10.3389/fnmol.2017.00083.

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.

Chapter 2
Exploring Polygenic Overlap Between ADHD and OCD

McKenzie Ritter and Yin Yao

Abstract Attention deficit hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) are neurodevelopmental disorders that onset during childhood. They are two of the most common psychiatric disorders affecting pediatric populations. ADHD and OCD share a common sub-phenotype, so it was thought that they may also share common risk alleles. Previously, ADHD and OCD had not been compared on a genome-wide association study (GWAS) or meta-analysis platform. Thus, a GWAS and meta-analysis were conducted for ADHD and OCD. The clinical overlap between the two disorders is also discussed in depth. Further research using larger sample sizes are warranted to increase power.

2.1 Introduction

Attention deficit hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) are pediatric neuropsychiatric disorders that commonly affect individuals worldwide. ADHD is characterized by recurring obsessions and/or compulsions, which affect approximately 5% of the population worldwide (Simon et al. 2009). ADHD is characterized by inattention, hyperactivity, and impulsivity (Polanczyk et al. 2007). OCD is characterized by recurring obsession and/or compulsions, where obsessions are unwanted thoughts, ideas, and impulses occurring more than once and compulsions are repetitive behaviors driven by obsessions (American Psychiatric Association 2016).

ADHD and OCD comorbidity has been recorded and found to range from 10% to 50% (Geller et al. 1996; Masi et al. 2006; Brem et al. 2014; Abramovitch et al. 2015). Previously recorded high comorbidity rates could have resulted because

M. Ritter · Y. Yao (✉)
Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA
e-mail: kay1yao@mail.nih.gov

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_2

ADHD often presents with features of inattention and distractibility, which ultimately could be misdiagnosed as OCD (Geller et al. 2002).

Several studies have been conducted that examined the overlapping subphenotypes between ADHD and OCD. Specifically, Sheppard et al. (2010) found that ADHD and OCD both have symptoms of inattention and distractibility that co-segregate in families. Additionally, Park et al. (2016) reported that hoarding, which is most commonly associated with OCD, may also be linked to executive functioning deficits in ADHD. It has also been noted that individuals with ADHD and OCD are thought to share diminished inhibitory control, conveyed as impulsivity in ADHD and poor control of obsessions and compulsions in OCD (Norman et al. 2016). Due to the previously reported clinical overlap between ADHD and OCD, the genetic link between the two disorders was also of interest.

A meta-analysis was performed between ADHD ($N = 3351$) and OCD ($N = 5415$). The ADHD sample consisted of 2064 trios, 896 cases, and 2455 controls. The OCD sample contained 2998 individuals from nuclear families. In addition to the meta-analysis, polygenic risk score (PRS) analyses were completed to test the hypothesis that multiple genes of small effect jointly contribute to the susceptibility of ADHD and OCD. An additional analysis was completed to examine protein-protein interactions to look for potential overlapping pathways between the two disorders using DAPPLE (http://www.broadinstitute.org/mpg/dapple/dapple.php). Then, an expression quantitative trait locus (eQTL) analysis was conducted to identify nonrandomly occurring genes associated with the prefrontal cortex region. The nominated genes from the eQTL and DAPPLE analysis were then used to explore the potential overlap between the two gene lists.

2.2 Polygenic Risk Score Analyses

Polygenic risk score (PRS) analyses summarize the genetic effects of a group of single-nucleotide polymorphisms (SNPs) that individually do not reach significance within an association study (Dudbridge 2013). The risk score was calculated as a sum of SNP alleles associated with a specific trait for an individual (Howie et al. 2009). The score was weighted by effect sizes estimated from a genome-wide association study (GWAS). These scores served to examine the genetic relationship between ADHD and OCD.

ADHD served as the discovery sample because the sample size was larger, and only summary statistics were available for this analysis. OCD was then used as the target dataset, and the PRSice (http://prsice.info/) software was used to calculate these scores. PRSice was designed to automate the steps of the PRS analyses by using both PLINK (http://zzz.bwh.harvard.edu//plink/) (Purcell et al. 2007) and R (https://www.r-project.org/) (R Core Team 2015). Linkage disequilibrium (LD) pruning was completed through PRSice using p-value thresholds of $p < 0.01$, 0.1, 0.2, 0.3, 0.4, and 0.5. Within each of the LD thresholds, p-value significance thresholds were determined, and R^2 values were calculated based on how well the regres-

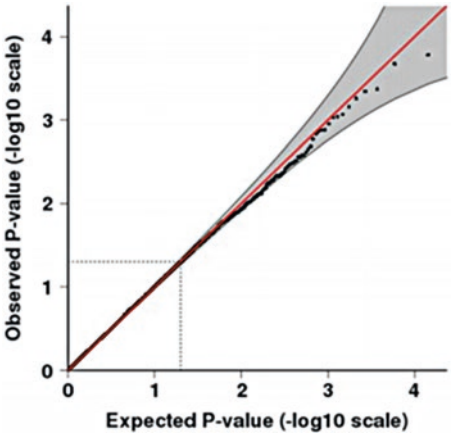


Fig. 2.1 Quantile-quantile (QQ) plot for *p*-values of the meta-analysis. QQ plots compare the observed vs. expected test statistic distributions. The shading indicates the 95% confidence intervals. The inflation factor λ is 1.008

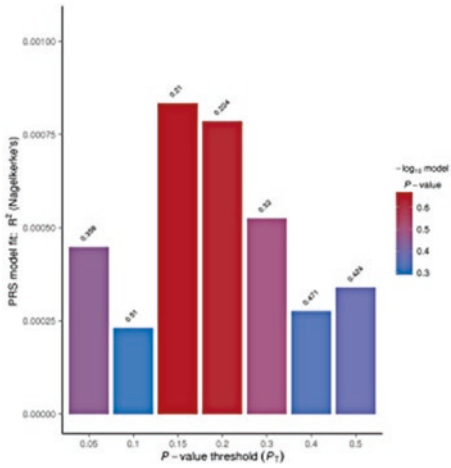


Fig. 2.2 PRSice bar plot for linkage disequilibrium (LD) threshold of 0.1. The tallest bar indicates the best fit polygenic risk score (PRS) for the ADHD PRS predicting OCD

sion fits the data (Fig. 2.1). For each of the *p*-value significance thresholds, quantitative polygenic scores were calculated for each individual within the target data. The scores were calculated by multiplying the number of risk alleles for each SNP (0, 1, or 2) by the score for that SNP, estimated from the discovery sample.

The created polygenic risk model was tested on the target sample to obtain the PRS for each individual. Logistic regression was then conducted to examine the relationship between risk score and the case-control status of the target data. PRSice automated the percentage of phenotypic variance that could be explained by the risk score (Fig. 2.2).

The R^2 value indicates how well the logistic regression approximates the data based on the *p*-value thresholds. The threshold of 0.15 had an R^2 value of 0.0834%. This means that approximately 0.08% of the data overlapped between the ADHD and OCD data.

2.3 Protein-Protein Link Evaluation

The DAPPLE (http://www.broadinstitute.org/mpg/dapple/dapple.php) software was used to conduct a protein-protein analysis to examine the connectivity between potential associated proteins (Rossin et al. 2011). DAPPLE seeks to find significant physical connectivity between proteins encoded by the genes found in the loci associated with the disease. The protein-protein interactions are based on reported biological information between proteins in InWeb, a database of 169,810 high-confidence pairwise interactions involving 12,793 proteins (Rossin et al. 2011). To test for the nonrandomness of the protein connections, DAPPLE was used to create random protein interaction networks with a within-degree node-label permutation method. Random networks each hold the same size, number of edges, and number of proteins with the same number of connections as the original network. Protein names in the random networks are randomly reassigned to proteins of equal protein connectivity, allowing for the evaluation of nonrandomness in the original network based on protein binding degree (Rossin et al. 2011).

The SNP's with a p-value <0.001 were included in the analysis to investigate whether any protein(s) associated with the disorders would give a statistically significant p-value. A total of 123 genes were included based on the before mentioned criterion. Six direct protein-protein interactions were identified and included ten total proteins (CHMP4B, EIF2S2, EIF3I, FGF10, FGFR2, ITCH, PIK3C2B, SELE, SELL, and UQCC; Fig. 2.3).

The direct connections from the analysis are shown in Table 2.1.

The overall direct connections protein interaction network had a p-value of 0.0879 (Fig. 2.3). Additionally, 543 indirect connections contributed to the network that linked the 6 direct protein interactions. None of the indirect connectors were of known biological relevance based on our current understanding of the diseases. A similar dapple analysis was conducted previously on an ADHD sample that resulted in no direct connections (Zayats et al. 2015).

2.4 Expression Quantitative Trait Locus (eQTL) Analysis

For the expression quantitative trait loci (eQTL) analysis, a p-value threshold of $p < 1.00 \times 10^{-4}$ was used in order to examine the relationship between the candidate SNPs and relevant eQTLs, using the eEQLAnalysis (http://hongbaocao.gousinfo.com/Software4Download.html) software. eEQLAnalysis can be used to conduct an eQTL analysis for the selected SNP list based on the BrainCloud (http://braincloud.jhmi.edu/) dataset (GSE30272). The BrainCloud dataset contains both SNP and gene expression data from 268 healthy subjects. The software includes three

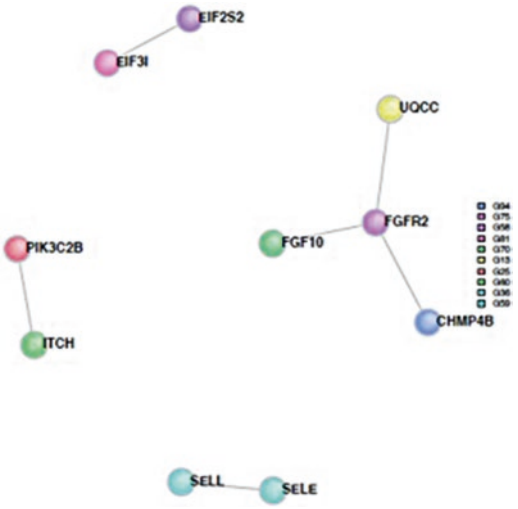


Fig. 2.3 Protein-protein interaction network built from proteins from the SNPs from the meta-analysis. The colored circles represent the proteins, while the different colors are associated with different regions. The gray lines represent the direct connections between the proteins

Table 2.1 Results of the protein-protein link evaluation in Disease Association Protein-Protein Link Evaluator (DAPPLE), direct connections

Protein	Region	Uncorrected *p*-value	Corrected *p*-value	Binding protein/s	Function
FGF10	G70	0.00199	0.00199	FGFR2, FGF10, UQCC	Fibroblast growth factor 10
FGFR2	G75	0.01394	0.01394	CHMP4B	Fibroblast growth factor receptor 2 ubiquinol-cytochrome C reductase complex
UQOC	G13	0.08598	0.08598	FGFR2	Assembly factor 1
EIF3I	G81	0.11629	0.11629	EIF2S2	Eukaryotic translation initiation factor 3
SELL	G59	0.11629	0.11629	SELL	Selectin E
EIF2S2	G58	0.16079	0.16079	EIF3I	Eukaryote translation initiation factor 2
SELL	G36	0.15529	0.15529	SELE	Selectin L
CHMP4B	G94	0.18442	0.18442	FGFR2	Charged multivesicular body protein 4A
PIK3C2B	G25	0.16444	0.16444	ITCH	Phosphatidylinositol-4-phosphate 3-kinase
ITCH	G60	0.24461	0.24461	PIK3C2B	Itchy E3 ubiquitin protein ligase

Table 2.2 Results of the expression quantitative trait locus (eQTL) analysis in eEQLAnalysis, top genes

Gene name	Chromosome number	eQTL *p*-value	Permutation *p*-value
LINC00314	21	1.461E-08	0.0039
CXCR2	2	5.934E-08	0.0029
ASB17	1	6.149E-07	0.0088
SELE	1	6.561E-07	0.0000
ACOT7	1	8.591E-07	0.0003
PRPS1L1	7	1.324E-06	0.0004
ZBF580	19	1.431E-06	0.0027
TAS2R41	7	1.663E-06	0.0049
ADAMTS20	12	1.710E-06	0.0027
WDFY3	4	2.136E-06	0.0041

modules: eQTL Map Generation, Permutation for Selected eQTLs, and report generation. The input requires a list of nominated SNPs. The outputs include the "significant SNPs" and associated statistics including eQTL *p*-values and permutation-based *p*-values. For more information about the software, please refer to http://hongbaocao.gousinfo.com/Software4Download.html.

SNPs with a *p*-value $<1.00 \times 10^{-4}$ from the genome-wide association tests were included in the eQTL analysis. This analysis was conducted in order to compare the results with the proteins identified from DAPPLE. The top SNPs associated with the prefrontal cortex are shown in Table 2.2.

2.5 Meta-analysis

The ADHD and OCD datasets were combined in order to conduct the meta-analysis. METAL (http://csg.sph.umich.edu/abecasis/Metal/) was used to conduct the analysis. Two thousand nine hundred ninety-eight OCD samples and 5415 ADHD samples were included in the analysis. A Manhattan plot and quantile-quantile (QQ) plot show the association *p*-values from the meta-analysis (Figs. 2.4 and 2.5, respectively).

2.6 Discussion

Psychiatric disorders, such as ADHD and OCD, are very complex and thus, clinically heterogenous. However, it has been previously reported that ADHD and OCD may potentially have overlapping sub-phenotypes. For example, Palumbo et al. (1997) suggested that ADHD, OCD, and autism have overlapping etiologies and

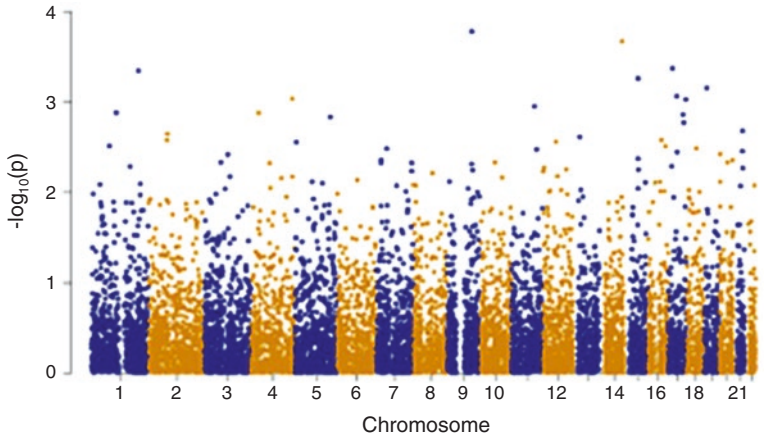


Fig. 2.4 Manhattan plot of all the genotyped and imputed SNPs for *p*-values of the meta-analysis between the attention deficit hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) studies

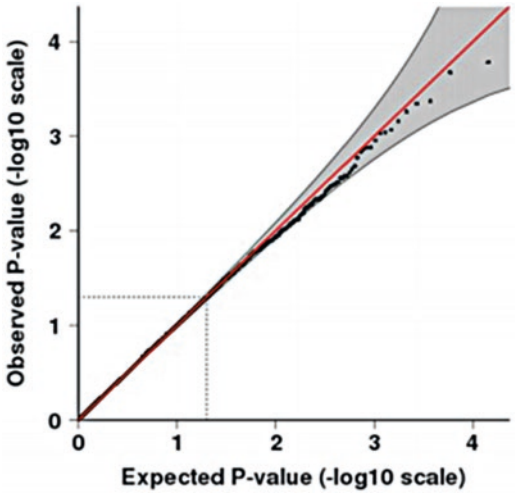


Fig. 2.5 Quantile-quantile (QQ) plot for *p*-values of the meta-analysis. QQ plots compare the observed vs. expected test statistic distributions. The shading indicates the 95% confidence intervals. The inflation factor λ is 1.008

thus are interrelated. In addition, Anholt et al. (2010) found that inattention plays a key role in obsessive-compulsive symptoms, which could further link ADHD and OCD. Even though the clinical overlap between ADHD and OCD has been reported, the genetic overlap found in this study was limited. It is possible that the association signals were diluted in the meta-analysis due to the heterogeneity of the samples.

Overall, a meta-analysis was conducted between ADHD and OCD. Though not significant, the SNP rs10989904 had the strongest association signal with a *p*-value of 1.65×10^{-4}. This SNP falls in an intergenic region but is located near the LOC100127962 pseudogene. None of the other SNPs hit in the analysis were of any

known biological relevance. In addition, the GWAS conducted on the ADHD Psychiatric Genomics Consortium data found the strongest signal on the CDH13 gene (Neale et al. 2010).

References

Abramovitch A, Dar R, Mittelman A, Wilhelm S. Comorbidity between attention deficit/hyperactivity disorder and obsessive-compulsive disorder across the lifespan. Harv Rev Psychiatry. 2015;23:245–62. https://doi.org/10.1097/HRP.0000000000000050.

American Psychiatric Association. What is obsessive-compulsive disorder? 2016. Available online at: https://www.psychiatry.org/patients-families/ocd/what-isobsessive-compulsive-disorder. Accessed 27 Apr 2016.

Anholt GE, Cath DC, Oppen PV, Eikelenboom M, Smit JH, van Megen H, et al. Autism and ADHD symptoms in patients with OCD: are they associated with specific OC symptom dimensions or OC symptom severity? Front Psych. 2010;40:580–9. https://doi.org/10.1007/s10803-009-0922-1.

Brem S, Grünblatt E, Drechsler R, Riederer P, Walitza S. The neurobiological link between OCD and ADHD. Atten Defic Hyperact Disord. 2014;6:175–202. https://doi.org/10.1007/s12402-014-0146-x.

Dudbridge F. Correction: power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9:4. https://doi.org/10.1371/annotation/b91ba224-10be-409d-93f4-7423d502cba0.

Geller DA, Biederman J, Griffin S, Jones J, Lefkowitz TR. Comorbidity of juvenile obsessive-compulsive disorder with disruptive behavior disorders. J Am Acad Child Adolesc Psychiatry. 1996;35:1637–46. https://doi.org/10.1097/00004583-199612000-00016.

Geller DA, Biederman J, Faraone SV, Cradock K, Hagermoser L, Zaman N, et al. Attention-deficit/hyperactivity disorder in children and adolescents with obsessive-compulsive disorder: fact or artifact? J Am Acad Child Adolesc Psychiatry. 2002;41:52–8. https://doi.org/10.1097/00004583-200201000-00011.

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:e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Masi G, Millepiedi S, Mucci M, Bertini N, Pfanner C, Arcangeli F. Comorbidity of obsessive-compulsive disorder and attention deficit/hyperactivity disorder in referred children and adolescents. Compr Psychiatry. 2006;47:42–7. https://doi.org/10.1016/j.comppsych.2005.04.008.

Neale BM, Medland SE, Ripke S, Asherson P, Franke B, Lesch KP, et al. Meta-analysis of genome-wide association studies of attentiondeficit/hyperactivity disorder. J Am Acad Child Adolesc Psychiatry. 2010;49:884–97. https://doi.org/10.1016/j.jaac.2010.06.008.

Norman LJ, Carlisi C, Lukito S, Hart H, Mataix-Cols D, Radua J, et al. Structural and functional brain abnormalities in attention-deficit/hyperactivity disorder and obsessive-compulsive disorder: a comparative meta-analysis. JAMA Psychiatry. 2016;73(8):815–25.

Palumbo D, Maughan A, Kurlan R. Hypothesis III: tourette syndrome is only one of several causes of a developmental basal ganglia syndrome. Arch Neurol. 1997;54:475–83. https://doi.org/10.1001/archneur.1997.00550160101023.

Park JM, Samuels JF, Grados MA, Riddle MA, Bienvenu OJ, Goes FS, et al. ADHD and executive functioning deficits in OCD youths who hoard. J Psychiatr Res. 2016;82:141–8. https://doi.org/10.1016/j.jpsychires.2016.07.024.

Polanczyk G, de Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: a systematic review and meta regression analysis. Am J Psychiatry. 2007;164:942–8. https://doi.org/10.1176/ajp.2007.164.6.942.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool-set for whole-genome association and populationbased linkage analysis. Am J Hum Genet. 2007;81:559–75. https://doi.org/10.1086/519795.

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immunemediated disease physically interact and suggest underlying biology. PLoS Genet. 2011;7:e1001273. https://doi.org/10.1371/journal.pgen.1001273.

Sheppard B, Chavira D, Azzam A, Grados MA, Umaña P, Garrido H, et al. ADHD prevalence and association with hoarding behaviors in childhood-onset OCD. Depress Anxiety. 2010;27:667–74. https://doi.org/10.1002/da.20691.

Simon V, Czobor P, Bálint S, Mészáros A, Bitter I. Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. Br J Psychiatry. 2009;194:204–11. https://doi.org/10.1192/bjp.bp.107.048827.

Zayats T, Athanasiu L, Sonderby I, Djurovic S, Westlye LT, Tamnes CK, et al. Genome-wide analysis of attention deficit hyperactivity disorder in Norway. PLoS One. 2015;10:e0122501. https://doi.org/10.1371/journal.pone.0122501.

Web Resources

http://www.broadinstitute.org/mpg/dapple/dapple.php
http://prsice.info/
http://zzz.bwh.harvard.edu//plink/
https://www.r-project.org/
http://hongbaocao.gousinfo.com/Software4Download.html
http://braincloud.jhmi.edu/
http://csg.sph.umich.edu/abecasis/Metal/

Chapter 3
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

3.1 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 disease risk factors. With increasing progress in gene identification, these

K. R. Merikangas (✉)
Genetic Epidemiology Research Branch, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
e-mail: Kathleen.merikangas@nih.gov

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_3

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.

3.1.1 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 (l) (Risch 1990b). Whereas l tends to exceed 20 for most autosomal dominant diseases, values of l 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).

3.1.2 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.

3.1.3 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.

3.1.4 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.

3.2 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).

3.2.1 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.

3.2.2 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.

3.2.3 Risk Estimation

Because genetic polymorphisms involved in complex diseases are likely to be non-deterministic (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-Eε4 (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 with 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 1291 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.

3.2.4 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).

3.3 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 (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 advances, it will be incumbent upon clinicians to become familiar with knowledge gleaned from genetic epidemiologic and genomics research. In the meanwhile, the 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. https://doi.org/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, ele-

vated 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. https://doi.org/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, JB MA, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, 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.
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. https://doi.org/10.1126/science.1091468.
Michael NL. Host genetic influences on HIV-1 pathogenesis. Curr Opin Immunol. 1999;11(4):466–74. https://doi.org/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. https://doi.org/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. https://doi.org/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):S10–6.
Varmus H. Getting ready for gene-based medicine. N Engl J Med. 2002;347(19):1526–7. https://doi.org/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. https://doi.org/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 4
Rare Variant 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 variant (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

4.1 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 (✉)
Case Western Reserve University, Cleveland, OH, USA

KeyBank, Cleveland, OH, USA
e-mail: tao.feng@case.edu

X. Zhu
Case Western Reserve University, Cleveland, OH, USA
e-mail: xzhu1@darwin.EPBI.CWRU.edu

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_4

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 variant analysis; in particular, we described the methods with great details in the context of next-generation sequencing data.

4.1.1 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 second quarter 2011, the US National Human Genome Resource Institute (NHGRI) GWAS catalogue lists 1449 genome-wide significant associations with 237 traits and diseases spread across all auto chromosome except the Y chromosome (Hindorff et al. 2011).

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), have 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 interac-

tions, such as gene-gene or gene-environment interactions (Bansal et al. 2010; Manolio et al. 2009; Zuk et al. 2012).

4.1.2 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.

4.2 Data Description and Methods

Below we first (in Sect. 4.2.1) 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. 4.2.2.1 and 4.2.2.11).

4.2.1 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 to denote the two alleles of jth SNP, in which A_i always refers to the rare allele and has an allele frequency pi. 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 $j = 1,\ldots, L$.

4.2.2 Methods

Between Sects. 4.2.2.1 and 4.2.2.11, we will provide mathematical details for each method illustrated and will also provide our insights of specific merits and limitations of each method.

4.2.2.1 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 ith 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).

4.2.2.2 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 to certain criteria (e.g., allele frequencies) and then collapse the rare variants within each group using the method described in Sect. 4.2.2.1. 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.

4.2.2.3 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 $w_j = \sqrt{Np_j\left(1-p_j\right)}$. 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.

4.2.2.4 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.

4.2.2.5 A Data-Adaptive Sum 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

$$\text{Logit}\Pr\left(y_i = 1\right) = \beta_{c0} + \sum_{j=1}^{L} x_{ij}\beta_c$$

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, $g\left(y_i\right) = \alpha_0 + \gamma\left[\sum_{j=1}^{L} w_j x_{ij}\right]$, where g is the link function and the weight w_j is define by $1/\sqrt{p_j\left(1-p_j\right)}$, similar as the Madsen and Browning's weight.

4.2.2.6 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 $n = 1$ is 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 that 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 jth variant, assume y_j is a binomial (n_j , p_j) if the rare variants were 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

$$T = \sum_{j=1}^{L}\left[\left(y_j - n_j p_0\right)^2 - n_j p_0\left(1-p_0\right)\right],$$

contrasts the variance of each observed count with the expected variance. The variance of T is derived by

$$c = \sum_{n=2}^{\max n} m(n)\sum_{u=0}^{n}\left[\left(u - np_0\right)^2 - np_0\left(1-p_0\right)\right]^2 f\left(u|,n,p_0\right),$$

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 ith 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 allows 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.

4.2.2.7 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

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z_i} + \boldsymbol{\beta}'\mathbf{X_i} + \varepsilon_i$$

when the phenotype are continuous traits and the logistic model

$$\text{logit}\,P\left(y_i = 1\right) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z_i} + \boldsymbol{\beta}'\mathbf{X_i},$$

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 βi 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 H0 :$\tau = 0$. To do the test, the variance-component score statistic

$$Q = \left(\mathbf{y} - \hat{u}\right)' \mathbf{K} \left(\mathbf{y} - \hat{u}\right)$$

is applied. In the above formula, $\mathbf{K} = \text{XWX'}$, $\hat{u}$ is the predicted mean of $\mathbf{y}$ under H_0, that is, $\hat{u} = \hat{\alpha}_0 \mathbf{Z}\hat{\alpha}$ for continuous traits and $\hat{u} = \text{logit}^{-1}\left(\hat{\alpha}_0 + \mathbf{Z}\hat{\alpha}\right)$ for dichotomous traits, and $\hat{\alpha}_0$ and $\hat{\alpha}$ 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 jth variant in ith individual, and $\mathbf{W} = \text{diag}(w_1, \ldots, w_L)$ contains the weights of the L variants. $\mathbf{K}$ is an $N \times N$ matrix with the (i,i')th element equal to $K\left(\mathbf{X}_i, \mathbf{X}_{i'}\right) = \sum_{j=1}^{L} w_j X_{ij} X_{i'j}$. $K(\bullet,\bullet)$ 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 $f(\mathbf{Xi})$ that 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 $K\left(\mathbf{X}_i, \mathbf{X}_{i'}\right) = \left(1 + \sum_{j=1}^{L} w_j X_{ij} X_{i'j}\right)^2$ or the weighted identity by state (IBS) kernel $K\left(\mathbf{X}_i, \mathbf{X}_{i'}\right) = \sum_{j=1}^{L} w_j \text{IBS}\left(X_{ij}, X_{i'j}\right)$. A question is how to choose w_j in the kernel

function, which can affect statistical power. Wu et al. (2011) suggested $\sqrt{w_j} \sim \text{Beta}\left(\text{MAF}_j;,a_1;,a_2\right)$, the beta distribution function with prespecified parameters a_1 and a_2 evaluated at the sample MAF using both cases and controls for the jth 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.

4.2.2.8 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 $P(y_i = 1) = \boldsymbol{\alpha}'\mathbf{Z_i} + \boldsymbol{\beta}'\mathbf{X_i}$ 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

$$\text{logit}\,\Pr\left(y_i = 1\right) = \tau S_i + \boldsymbol{\gamma}'\mathbf{Z_i},$$

where $S_i = \boldsymbol{\xi}'\mathbf{X_i}$. Note that $\xi = (\xi_1 ,\ldots, \xi_L)'$ is a $L \times$ vector of weights and that S_i is a weighted linear combination of x_{i1} , x_{i2} ,… 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

$$U = \sum_{i=1}^{N}\left(y_i - \frac{e^{\hat{\gamma}'\mathrm{Z_i}}}{1+e^{\hat{\gamma}'\mathrm{Z_i}}}\right)s_i,$$

where $\hat{\gamma}$ is the restricted maximum likelihood estimator of γ and solves the equation $\sum_{i=1}^{N}\left(y_i - \frac{e^{\hat{\gamma}'\mathrm{Z_i}}}{1+e^{\hat{\gamma}'\mathrm{Z_i}}}\right)Z_i = 0$. 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 $v_i = \dfrac{e^{\hat{\gamma}' Z_i}}{\left(1 + e^{\hat{\gamma}' Z_i}\right)^2}$. 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 $\bar{y} = N^{-1} \sum_{i=1}^{N} y_i$.

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 $U_k = \sum_{i=1}^{N} \left(Y_i - \dfrac{e^{\hat{\gamma}' Z_i}}{1 + e^{\hat{\gamma}' Z_i}} \right) S_{ki}$ 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 $U_{kl} = \left(y_i - \dfrac{e^{\hat{\gamma} Z_i}}{1 + e^{\hat{\gamma} Z_i}} \right) \left\{ S_{ki} - \left(\sum_{i=1}^{N} v_i S_{ki} Z_i \right) \left(\sum_{i=1}^{N} v_i Z_i Z_i' \right)^{-1} Z_i \right\}$.

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 $x_j = 0$ 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 $Q = \sum_{j=1}^{L} \xi_j U_j^2$, here 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.

4.2.2.9 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, assume the 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 $p0$ in the controls. Then the risk haplotype set is defined as

$$S = \left\{ h_i \left| p_i - p_i^0 > \gamma \sqrt{\frac{p_i^0 \left(1 - p_i^0\right)}{2n}} \right. \right\},$$

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, and then the rare risk haplotype set for affected sibpairs can be defined by

$$S = \left\{ h_i \left| p_i - p_i^0 > \gamma \sqrt{\frac{p_i^0 \left(1 - p_i^0\right)}{3M}} \right. \right\}$$

where h, p, and p^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 haplotype 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 add 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.

4.2.2.10 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 jth 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 observation, 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 jth 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 x 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 $\dfrac{\hat{\gamma}_j}{sd}$, 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

$$w_j = \begin{cases} \gamma_j, & \text{if } \gamma_j > \bar{\gamma} + c\sigma \text{ or } \gamma_j < \bar{\gamma} - c\sigma \\ 0, & \text{otherwise} \end{cases},$$

where σ is the standard deviation calculated from γ_j, $j = 1,\ldots L$, $c = 1.64$ or 1.28 is a parameter, and $\bar{\gamma}$ 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.

4.2.2.11 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 ith sibpair's genotypes of the L SNPs as $g_i = ((g_{i11}, g_{i21}), (g_{i12}, g_{i22}), \ldots, (g_{i1L}, g_{i2L}))$ where (g_{i1j}, g_{i2j}) refers to the jth SNP's genotypes for the ith sibpair. There is no need to differentiate the first or second sib here. The idea here is that if A at the jth 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 jth SNP. To do this, two scenarios are considered. First, if both affected sibs carry A at the jth SNP, the weight of A at this SNP is proportional to

$$\frac{\Pr\left(\text{both sibs are affected} \middle| \left(g_{i1j}, g_{i2j}\right) = (A\sim, A\sim)\right)}{p\left(\text{both sibs are affected}\right)} = \frac{\Pr\left(\left(g_{i1j}, g_{i2j}\right) = (A\sim, A\sim) \middle| \text{both sibs are affected}\right)}{\phi_1}$$

where $\phi_1 = Pr\,((g_{i1j}, g_{i2j}) = (A\sim, A\sim))$. 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

$$\frac{\Pr\left[\text{both sibs are affected} \middle| \left(g_{i1j}, g_{i2j}\right) = (A\sim, aa), A \text{ present at other sites of sib 2}\right]}{p\left(\text{both sibs are affected}\right)} = \frac{\Pr\left[\left(g_{i1j}, g_{i2j}\right) = (A\sim, aa), A \text{ present at other sites of sib 2} \middle| \text{both sibs are affected}\right]}{\phi_2}$$

where $\phi_2 = P[(g_{i1j}, g_{i2j}) = (A\sim, aa), A$ present at other sites of sib 2]. 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

$$\tilde{g}_{ij} = \begin{cases} \dfrac{1}{\phi_2}, & \text{when}\left(g_{i1j}, g_{i2j}\right) = \left(A\sim, \text{A}\sim\right) \\ \dfrac{L_0}{2L\phi_2}, & \text{when}\left(g_{i1j}, g_{i2j}\right) = \left(A\sim, aa\right), \text{and} L_0 \text{of the other } SNPs \text{ carry an } A \\ & \text{allele for sib 2} \\ 0, & \text{otherwise} \end{cases}.$$

In the above equation, the second term was divided by two 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

$$\begin{aligned} \phi_1 &= \Pr\left(\left(g_{i1j}, g_{i2j}\right) = (A\sim, A\sim)\right) \\ &= \sum_{I=1}^{2} \Pr\left[\left(g_{i1j}, g_{i2j}\right) = (A\sim, A\sim) \middle| I\right] \\ &= p_j\left(1 + \frac{1}{4}p_j - \frac{1}{4}p_j^2\right) + \frac{1}{4}p_j^2\left[1 + 2p_j\left(1-p_j\right) + 3\left(1-p_j\right)^2\right] \end{aligned}$$

and

$$\begin{aligned} \phi_2 &= \Pr\left[\left(g_{i1j}, g_{i2j}\right) = (A\sim, aa), A \text{present at other sites of sib 2}\right] \\ &= \Pr\left[\left(g_{i1j}, g_{i2j}\right) = (A\sim, aa)\right]\left(1 - P\left[A \text{ not present at any sites}\right]\right) \\ &= \sum_{I=1}^{2} \Pr\left[\left(g_{i1j}, g_{i2j}\right) = (A\sim, aa) \middle| I\right] P(I)\left(1 - P\left[A \text{ not present at any sites}\right]\right) \\ &= \frac{1}{4}p_j\left(1-p_j\right)^2\left(4-p_j\right)\left[1 - \prod_{k\neq j}^{L}\left(1-p_k\right)^2\right] \end{aligned}$$

where p_j is the A allele frequency at the jth 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 jth SNP, we then calculate $\gamma_j = \dfrac{1}{N_{\text{sib}}}\sum_{i=1}^{N_{\text{sib}}}\tilde{g}_{ij}$, 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 jth SNP to be

$$w_j = \begin{cases} \gamma_j, & \text{if } \gamma_j > \overline{\gamma} + c\sigma \\ 0, & \text{otherwise} \end{cases},$$

where $\bar{\gamma}$ and σ are the mean and standard deviation calculated from γ_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 jth SNP should be proportional to

$$\frac{P\left(\text{sibs 1 are affected and sib 2 is not}\middle|\left(g_{i1j},g_{i2j}\right)=(A\sim,aa)\right)}{p\left(\text{sibs 1 are affected and sib 2 is not}\right)}$$
$$=\frac{P\left(\left(g_{i1j},g_{i2j}\right)=(A\sim,aa)\middle|\text{sibs 1 are affected and sib 2 is not}\right)}{\phi_3},$$

where $\phi_3 = P\left(\left(g_{i1j},g_{i2j}\right)=\left(A\sim aa\right)\right)=\frac{1}{4}p_j\left(1-p_j\right)^2\left(4-p_j\right)$. For the ith discordant sibpair and the jth SNP, the genotype score is

$$\tilde{g}_{ij}\begin{cases}\frac{1}{\phi_3}, & \text{when}\left(g_{i1j},g_{i2j}\right)=\left(A\sim,aa\right)\\ 0, & \text{otherwise}\end{cases}.$$

In the same way as for affected sibpairs, the weights for discordant sibpairs can be defined.

4.3 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, MI MC, 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):1810–5.

Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3000 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, Hall PN, Klemm AK, Manolio TA. A catalog of published genome-wide association studies. 2011. Available at: www.genome.gov/gwastudies. Accessed 15 Sept 2012.

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010;5(11):e13584. https://doi.org/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(a) 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 5
Whole-Genome Association Analysis of Treatment Response from Obsessive-Compulsive Disorder

McKenzie Ritter and Haide Qin

Abstract Up to 30% of individuals with obsessive-compulsive disorder (OCD) display an inadequate response to serotonin reuptake inhibitors (SRIs). Genetic predictors of OCD treatment response have not been efficiently examined using a genome-wide association study (GWAS). In order to identify genetic variations that could potentially influence SRI response, a GWAS with 804 OCD patients containing information on SRI response was conducted. SRI response was used based on self-reported data and characterized as "response" (N = 514) or "non-response" (N = 290). The more powerful quasi-likelihood score (MQLS) test was used to conduct a genome-wide association test correcting for relatedness. An adjusted logistic model was then used to examine the effect size of the variants in probands. The most significant SNP found was rs17162912 ($P = 1.76 \times 10^{-8}$), which is near the gene *DISP1* on 1q41–q42, a microdeletion region that has been implicated in neurological development. Six other SNPs showed evidence of association ($P < 10^{-5}$): rs9303380, rs12437601, rs16988159, rs7676822, rs1911877, and rs723815. Two of the SNPs in strong linkage disequilibrium, rs7676822 and rs1911877, are located near the *PCDH10* gene and had p-values of 2.86×10^{-6} and 8.41×10^{-6}, respectively. The 35 other variations with a p-value $<10^{-4}$ are involved with multiple genes expressed in the brain, including *BRIN2B*, *PCDH10*, and *GPC6*. The enrichment analysis suggested that there may be genes that play a role in the glutamatergic neurotransmission system (FDR = 0.0097) and the serotonergic system (FDR = 0.0213). The results of this study could provide new insights into genetic mechanisms underlying treatment response in OCD, but studies with larger sample sizes and more detailed information on drug dosage, as well as treatment duration, are needed.

M. Ritter
Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA

H. Qin (✉)
Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, People's Republic of China

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_5

5.1 Introduction

Approximately 1–3% of the US population has obsessive-compulsive disorder (OCD), which is a neuropsychiatric disorder characterized by recurrent obsession and/or compulsions that cause distress and impairment (American Psychiatric Association 1994). OCD tends to aggregate in families, and segregation analyses and twin studies support a genetic influence (Nestadt et al. 2010). A genome-wide linkage study found several OCD susceptibility loci (3q, 7p, 1q, 15q, and 6q) (Shugart et al. 2006). Several variants were previously found to be associated with OCD, including *SLC1A1* (Shugart et al. 2009), *SLC6A* (Voyiaziakis et al. 2012; Murphy and Lesch 2008), and *GRIN2B* as potential loci (Arnold et al. 2004; Stewart et al. 2007; Alonso et al. 2012). Previously, association studies found a set of candidate genes that were inconsistently reported to be associated with OCD (Taylor 2013). As of recent, the shift in genome-wide association studies revealed genes *PTPRD*, *DLGAP1*, *CDH10*, and *GRIK2* as potential OCD susceptible loci (Stewart et al. 2013; Mattheisen et al. 2014).

Typical treatment for OCD includes a combination of exposure response prevention (ERP) and medication. Serotonin reuptake inhibitors (SRIs) are most commonly used for OCD. Unfortunately, up to 30% of patients treated with SRIs show either poor or no treatment response to the medication or cannot handle the adverse effects of the SRIs (Ferguson. 2001). The current literature on genetic predictors of SRI treatment response in OCD is scarce (Di Bella et al. 2002; Brandl et al. 2012, 2014). Therefore, further research on the genetic variants influencing treatment response is warranted.

SRIs inhibit the reuptake of serotonin by presynaptic cells, thus increasing extracellular levels of serotonin in the synaptic cleft, allowing serotonin to more easily bind to the postsynaptic receptor (Murphy and Lesch 2008; Sangkuhl et al. 2009). Over 60 proteins are known to play a role in the serotonin signaling pathway. Among these proteins, the serotonin transporter gene, *SLC6A*, could impact SRI response (Murphy and Lesch 2008). Additionally, genetic variants in other genes (*CYP2D6*, *SLC1A1*, *SLC6A4*, *HTR1B* receptor, *5-HT2A* receptor, and *BDNF*) have been reported to influence SRI response in OCD (Brandl et al. 2012). Several of these studies had small sample sizes and a limited number of known genetic variations in candidate genes. In addition, analytical approaches vary widely among different studies, which could have led to the inconsistency of the results. For example, in 2012, Tansey et al. reported results from the first GWAS of SRI response in major depression (2012). All of the findings reported in this chapter were based on a previously published manuscript (Qin et al. 2015). Therefore, an important research question is whether genetic variations influence SRI treatment response in OCD. To address this question, a whole-genome association analysis was performed on the response to SRIs in 804 OCD cases using a novel, more powerful quasi-likelihood score (MQLS) test to correct for relatedness. This chapter is based on a previously published manuscript (Qin et al. 2015).

5.2 Material and Methods

The samples used in this study were originally recruited as a part of the OCD Collaborative Genetics Association Study (OCGAS). The detailed methods of recruitment have been previously recorded (Samuels et al. 2006; Nestadt et al. 2010). A detailed description of the data and scales used for diagnosis are documented in the complete manuscript (Qin et al. 2015).

Treatment response was dichotomized into "response" and "non-response" for the analyses. Genotyping and quality control measures were then completed, which can be found in the complete manuscript as well (Qin et al. 2015).

5.3 Statistical Methods

The MQLS test was conducted to complete association tests and correct for their relatedness coefficients (based on identity-by-descent). A logistic model adjusted for sex and age was used to evaluate effect size in the probands using PLINK (Purcell et al. 2007). The association test has the potential to underestimate true signals of association with SRI response due to limited statistical power, all variations with MQLS test p-values $<10^{-4}$ with a large effect size (odds ratio ≥ 1.50 for the risk allele) were reported. All statistical analyses were conducted using in-house R scripts on a GentOS-based cluster computer.

SNP annotation was completed using SNP-NEXUS based on dbSNP135/hg19 (Chelala et al. 2009). Cross references to other GWAS association studies were explored using the NHGRI GWAS catalogue (Hindorff et al. 2013). Neurobiological evidence was also examined through peer-reviewed publications in the PubMed database. Additionally, LD plots were completed using the LocusZoom software based on 1000 genome CEU population data (hg19/1000 Genomes Mar 2012 EUR) (Pruim et al. 2010).

5.4 Results

Table 5.1 lists the demographic and clinical characteristics of the samples. After quality control of the data, 5597,847 SNPs (81.8% of the SNPs attempted in the array) were genotyped successfully. A total of 804 individuals with informative drug effect data (514 responders and 209 non-responders) had a call rate of 99.9%. Figure 5.1a shows a QQ plot. Of the 42 SNPs with a p-value $<10^{-4}$, one SNP met the genome-wide significance level ($p = 1.76 \times 10^{-8}$) for SRI treatment response. Additionally, six SNPS showed suggestive evidence of association with a p-value $<10^{-5}$, and 35 SNPs showed signals of association at the level of $p < 10^{-4}$ (Fig. 5.1b and Table 5.2).

Table 5.1 Characteristics of OCD participants

Group	Subgroup	Count ($N = 804$)	Frequency
Sex			
	Male	312	0.39
	Female	492	0.61
Age[a]			
	7–9	19	0.02
	10–19	189	0.23
	20–29	170	0.21
	30–39	173	0.22
	40–49	159	0.20
	50–78	94	0.12
Age at onset of OC symptoms			
	5–9	518	0.64
	10–19	237	0.30
	20–44	118	0.06
SRI response[b]			
	"No response"	290	0.36
	"Response"	514	0.64

[a]Age unknown for five participants
[b]"Couldn't tolerate" and "unknown" were excluded from data analysis

The most significant SNP, rs17162912, is located in proximity (distance of ~13 kb) to the *DISP1* gene ($p = 1.76 \times 10^{-8}$; OR = 0.39 [95% CI 0.26–0.58] (Table 5.2 and Fig. 5.2a left panel). Since there were no nearby markers with complete LD with rs17162912, genotypes in the left and right regions flanking that SNP were imputed (up to 250 kb), as well as an association test. The results showed that SNPs in strong LD with rs17162912 also presented suggestive association signals (Fig. 5.2a right panel). The integrated ENCODE regulation databases were explored, and it was found that rs17162912 is close to a peak (approximately 13 kb) of the H3K27AC protein-binding score, suggesting that this region encompasses the promoter of *DISP1*. *DISP1* encodes a 12-transmembrane domain protein that is required for long-range sonic hedgehog (Shh) secretion and transport. This is important in the establishment of cell-cell contact and spinal cord development (Etheridge et al. 2010).

rs7676822 and rs1911877 were SNPs with suggestive signals, which are located near the *PCDH10* gene (distance = 1818 kb and 1772 kb, respectively), and showed *p*-values of 2.86×10^{-6} (OR = 0.65 [95% CI o.51–0.83]) and 8.41×10^{-6} (OR = 0.66 [95% CI 0.52–0.84]), respectively (Fig. 5.2b). The LD relationship between rs7676822 and rs1911877 means that the two SNPs should be considered as one hit. It is worth mentioning that *PCDH10* belongs to a protocadherin gene family that consists of the largest subgroup of the cadherin superfamily, which mediates cell-cell adhesion and intracellular signaling. Most PCDHs (protocadherins) are predominantly expressed in the central nervous system and have been suggested to play crucial roles in both the formation and maintenance of synaptic functioning (Kim et al. 2007).

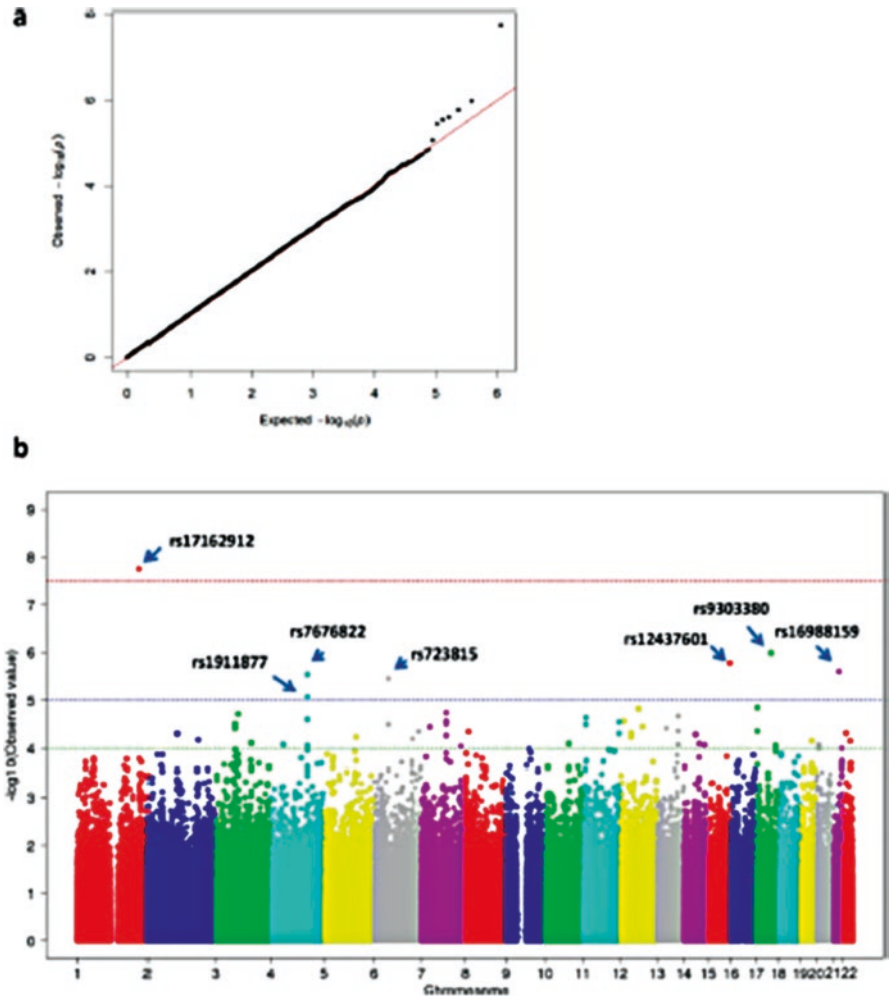


Fig. 5.1 Genome-wide association study of genetic variations and treatment response. (**a**) Q-Q plot for the association test of genetic variations. (**b**) Manhattan plot for the association test of genetic variations and SRI response. MQLS test was performed to test the association of variants associated with drug response. A red line indicates genome-wide significance (5×10^{-8}); a blue line indicates the level of suggestive evidence for association (1×10^{-5})

To comprehensively evaluate the role of some known pathways in the nervous system, an enrichment analysis was conducted. This was done to test whether there were any genes that were significantly enriched in various neuron signaling pathways. It was found that both the glutamatergic neurotransmission pathway and serotonergic neurotransmission pathways displayed more than a twofold enrichment. The glutamatergic pathway had the highest enrichment score of 3.38, as well as the best false discovery rate (FDR) of 0.0097. The serotonergic pathway had the second best enrichment score of 2.39, with an FDR or 0.0213 (Table 5.3 and Fig. S1).

Table 5.2 Top loci associated with treatment response in OCD patients

SNP	Chr.	position	A1/A2[a]	Resp.[b]	Non-resp.[b]	P[c]	OR(95%CI)[d]	Region	Nearest gene (distance/bp)
rs17162912	1	222974926	C/T	0.06	0.15	1.76 × 10–8	0.39(0.26–0.58)	Intergenic	DISP1(13505)
rs9303380	17	54117492	A/G	0.03	0.07	1.03 × 10–6	0.37(0.21–0.64)	Intergenic	ANKFN1(113344)
rs12437601	15	98687330	C/T	0.09	0.03	1.66 × 10–6	4.07(2.16–7.66)	Intergenic	ARRDC4(170262)
rs16988159	21	32727653	C/T	0.3	0.42	2.48 × 10–6	0.57(0.45–0.73)	Intronic	TIAM1
rs7676822	4	132252355	G/T	0.28	0.39	2.86 × 10–6	0.65(0.51–0.83)	Intergenic	PCDH10(1818115)
rs723815	6	52519203	A/C	0.2	0.11	3.50 × 10–6	2.06(1.46–2.9)	Intergenic	LOC730101(9996)
rs1911877	4	132298239	C/T	0.3	0.4	8.41 × 10–6	0.66(0.52–0.84)	Intergenic	PCDH10(1772231)
rs8081611	17	4813365	C/T	0.12	0.05	1.40 × 10–5	2.59(1.6–4.19)	Intergenic	CHRNE(6996)
rs7972963	12	66646199	T/G	0.08	0.14	1.50 × 10–5	0.54(0.37–0.78)	UTR3	IRAK3
rs17253738	13	94874089	A/G	0.14	0.21	2.13 × 10–5	0.59(0.43–0.82)	Intronic	GPC6
rs2706652	11	12289058	A/G	0.42	0.33	2.30 × 10–5	1.5(1.18–1.92)	Intergenic	MICAL2(3727)
rs7972211	12	14269986	G/A	0.16	0.23	2.71 × 10–5	0.65(0.49–0.87)	Intergenic	GRIN2B(136964)
rs318982	11	131415267	T/C	0.21	0.29	2.82 × 10–5	0.65(0.5–0.86)	Intronic	NTM
rs6918918	6	52515078	T/C	0.22	0.13	3.17 × 10–5	1.94(1.4–2.68)	Intergenic	LOC730101(14121)
rs11022029	11	11806317	C/T	0.13	0.2	3.18 × 10–5	0.65(0.48–0.87)	Intergenic	USP47(56653)
rs881499	7	30976064	C/T	0.25	0.36	3.58 × 10–5	0.55(0.42–0.72)	Intergenic	AQP1(10933)
rs905690	3	68725295	T/C	0.35	0.26	3.79 × 10–5	1.56(1.2–2.02)	Intergenic	FAM19A4(55620)
rs12561532	13	52108978	G/A	0.06	0.12	3.81 × 10–5	0.48(0.32–0.72)	Intergenic	MIR4703(17747)
rs9516369	13	94868584	G/A	0.14	0.21	4.38 × 10–5	0.61(0.44–0.84)	Intronic	GPC6
rs7214776	17	4811615	C/T	0.12	0.06	4.39 × 10–5	2.4(1.51–3.83)	Intergenic	CHRNE(5246)
rs9365319	6	162114707	T/C	0.13	0.21	4.49 × 10–5	0.57(0.42–0.77)	Intronic	PARK2
rs7004833	8	11840011	G/A	0.05	0.1	4.53 × 10–5	0.47(0.3–0.75)	Intronic	DEFB135
rs4768165	12	40025034	A/G	0.25	0.34	4.79 × 10–5	0.66(0.51–0.85)	Intronic	C12orf40
rs6005451	22	27852183	C/T	0.09	0.16	4.85 × 10–5	0.53(0.37–0.75)	Intergenic	MN1(292082)
rs10894396	11	131326035	A/G	0.41	0.29	4.91 × 10–5	1.72(1.34–2.22)	Intronic	NTM

rs2293223	2	103035468	T/C	0.15	0.24	4.92 × 10–5	0.6(0.44–0.8)	Intronic	IL18RAP
rs1403552	2	103088777	A/G	0.15	0.24	5.00 × 10–5	0.59(0.44–0.79)	Upstream	SLC9A4
rs11158347	14	61930678	A/G	0.33	0.21	5.18 × 10–5	1.83(1.39–2.41)	Intronic	PRKCH
rs7706447	5	116513164	C/A	0.04	0.1	5.83 × 10–5	0.36(0.23–0.59)	Intergenic	LOC728342(238044)
rs11611119	12	40166257	C/T	0.35	0.26	5.83 × 10–5	1.59(1.22–2.07)	Intronic	SLC2A13
rs4596498	6	139540103	A/G	0.24	0.16	6.36 × 10–5	1.73(1.26–2.36)	Intergenic	TXLNB(21096)
rs7565966	2	179742232	C/T	0.45	0.33	6.69 × 10–5	1.63(1.28–2.08)	Intronic	CCDC141
rs12974044	19	42368629	G/A	0.37	0.27	7.07 × 10–5	1.56(1.2–2.02)	Intronic	RPS19
rs139531	22	41676176	G/A	0.3	0.2	7.07 × 10–5	1.69(1.28–2.24)	Intronic	RANGAP1
rs1471659	3	126812577	G/A	0.11	0.17	7.74 × 10–5	0.61(0.43–0.85)	Intergenic	PLXNA1(56342)
rs4933958	10	85821027	C/T	0.29	0.21	8.04 × 10–5	1.56(1.18–2.05)	Intergenic	GHITM(78158)
rs10013818	4	44293409	T/C	0.26	0.18	8.34 × 10–5	1.6(1.19–2.15)	Intronic	KCTD8
rs3891616	13	94866849	C/A	0.14	0.2	8.39×10^{-5}	0.63(0.46–0.87)	Intronic	GPC6
rs722665	20	8508604	C/T	0.4	0.29	8.47×10^{-5}	1.61(1.25–2.08)	Intronic	PLCB1
rs2295394	14	93412743	P/C	0.04	0.08	8.55×10^{-5}	0.48(0.29–0.79)	NA	NA
rs351098	4	132409029	T/C	0.22	0.3	8.77×10^{-5}	0.67(0.51–0.86)	Intergenic	PCDH10(1661441)
rs12532545	7	141875267	A/C	0.17	0.25	9.21 × 10–5	0.63(0.48–0.84)	Intronic	LOC100124692

Abbreviations: *Chr* chromosome number, *A1/A2 OR* odds ratio, *CI* confidence interval, *MQLS* a more powerful quasi-likelihood score test

[a]A1/A2, in which "A1" is minor allele and "A2" is major allele

[b]Resp. minor allele sequence (MAF) for the patient, response to SSRIs; non-resp., MAF for the patients non-response to SSRIs

[c]MQLS_*robust p*-value, cutoff *p*-value threshold was set 1×10^{-4} for the risk allele

[d]Logistic regression model was performed on probands, adjusted by sex and age. Cutoff threshold was set at OR ≥ 1.5 for the risk allele

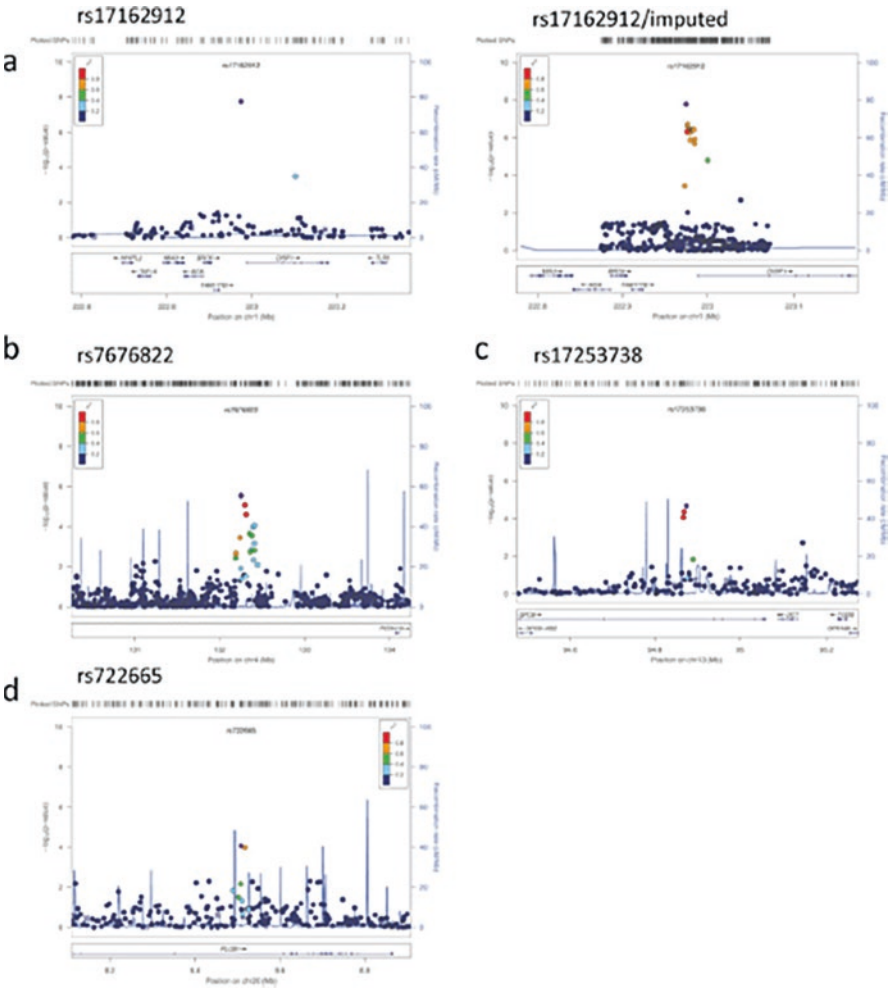


Fig. 5.2 Regional association plot with LD illustrated for significant SNPs. (**a**) SNAP plot of rs17162912 for the association test (left) and for the association test after the imputed SNPs was included (right). (**b**) SNAP plot of rs7676822, rs17253738 (**c**), and rs722665 (**d**)

The glutamatergic pathway contained the SNP rs7972211 near the *GRIN2B* gene, which is an important component of the glutamatergic neurotransmission system. This SNP showed a signal of association with SRI response with a *p*-value of 2.71×10^{-5} (OR = 0.65 [95% CI 0.49–0.87]) (Table 5.2). Additionally, *GPC6*, another gene of the glutamatergic transmission system, had three SNPs (rs17253738, rs9516369, rs3891616) that exhibited association signals with $p = 2.13 \times 10^{-5}$ (OR = 0.59 [95% CI 0.43–0.82]), $p = 4.38 \times 10^{-5}$ (OR = 0.61 [95% CI 0.44–0.84]), and $p = 8.39 \times 10^{-5}$ (OR = 0.63 [95% CI 0.46–0.87]), respectively (Table 5.2 and Fig. 5.2c). There is tight LD that exists among these three SNPs and thus serves as

Table 5.3 Enrichment analysis results in ten neurologically relevant pathways

Pathways examined	Genes enriched	Enrichment score	*P*-value	FDR
Glutamatergic signaling	14	3.38	0.0009	0.0097
Serotonergic signaling	11	2.39	0.0047	0.0213
Long-term potentiation	6	1.55	0.0058	0.0213
Neurotrophin signaling pathway	8	1.54	0.0120	0.0330
Long-term depression	4	1.04	0.0280	0.0512
GABAergic signaling	7	1.12	0.0340	0.0512
Dopaminergic synapse	7	1.02	0.0346	0.0511
Retrograde endocannabinoid signaling	6	0.88	0.0372	0.0512
Cholinergic signaling	4	0.67	0.5720	0.6292
Synaptic vesicle cycle	3	0.34	0.9280	0.9281

one hit. *GPC6* is known to promote the glutamate receptor clustering and receptivity and also induces the formation of postsynaptic signaling in the synapses of the central nervous system. Exhaustion of *GPC6* significantly reduces its function in inducing postsynaptic activity (Allen et al. 2012). Interestingly, both *DLGAP1* and *DLGAP2* support the enrichment (Fig. S1a). *DLGAP1* has recently been suggested as an OCD susceptibility gene (Stewart et al. 2013).

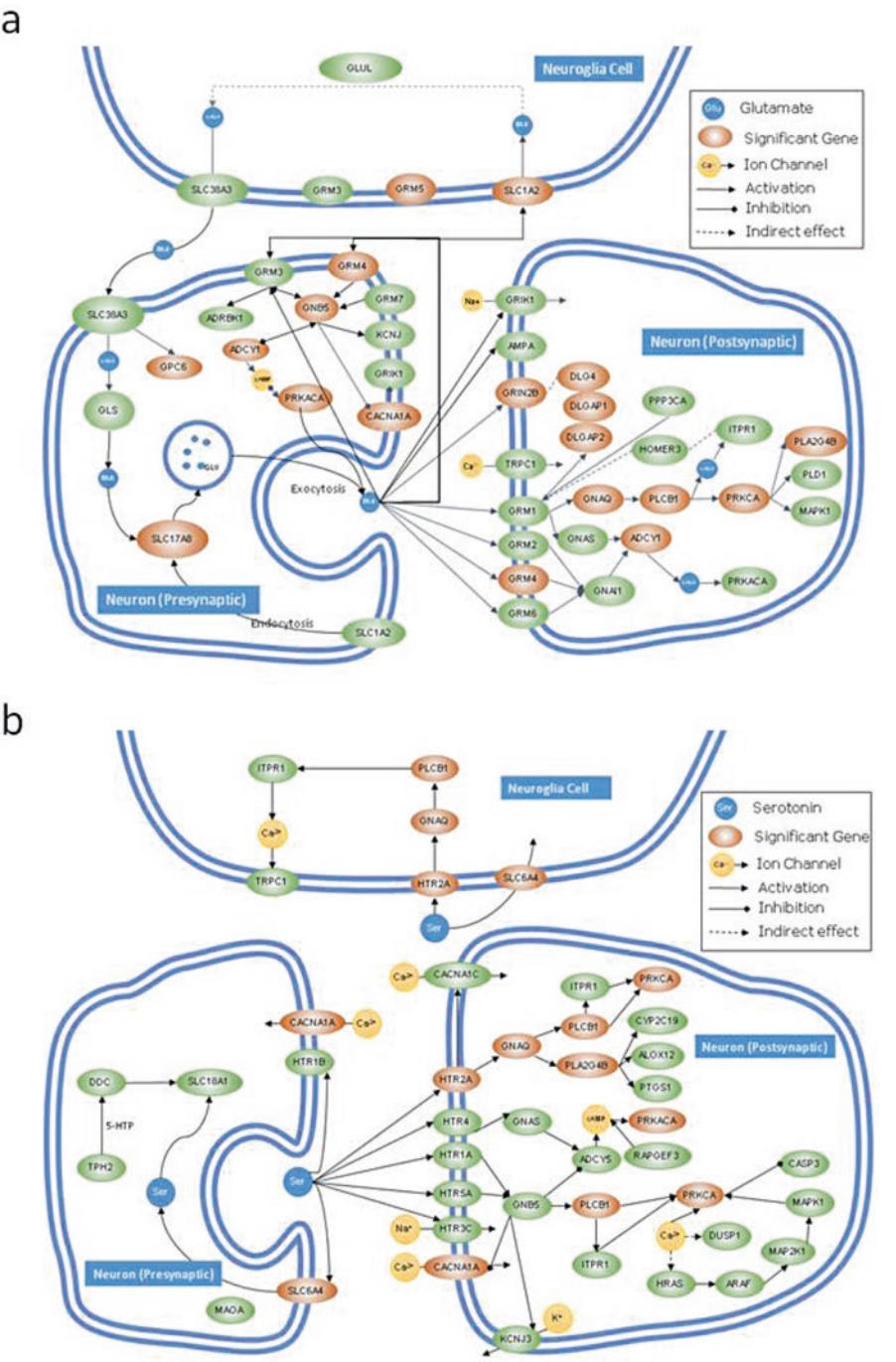
Two within LD ($R^2 = 0.6$) variants (rs722665 and rs2423366) of the serotonergic system in the *PLCB1* gene showed an association at $p = 8.47 \times 10^{-5}$ (OR = 1.83 [95% CI 1.39–2.41]) (Table 5.2 and Fig. 5.2d). Several other well-established genes including *HTR2A* and *SLC6A4* appeared to support the enrichment (Fig. S1b).

5.5 Discussion

The association between genetic variations and treatment response in OCD was tested in this study. Replication is warranted, but this study adds to the comprehension of how low genetic variants could contribute to drug response in OCD treatment. The top SNP hit from the GWAS was rs17162912, which is located near the *DISP1* gene. Additionally, the enrichment analysis indicated roles of genes in the glutamatergic neurotransmission system (FDR = 0.0097) and the serotonergic system (FDR = 0.0213).

The *DISP1* gene is located in the 1q41–q42 locus, which has a microdeletion related to a syndrome with symptoms of significant mental retardation, behavioral problems, seizures, as well as characteristic dysmorphic features (Jun et al. 2013). Even though rs17162912 does not fall within gene regulators, it is within close proximity to a promoter of the *DISP1* gene in ENCODE databases.

Another gene that plays a role in cell-cell contact, *PCDH10*, is an autism spectrum disorder (ASD)-related gene (Morrow et al. 2008). This gene had a suggestive level of association with SRI response. Other GWAS have shown that several PCDH genes are associated with various neuropsychiatric disorders, including autism,



Supplementary Figure S1 Illustration of genes enriched in glutamatergic system and serotonergic system. (**a**) The serotonin signaling pathway. (**b**) The glutamate signaling pathway. Note: the ovals indicate the molecules in the synaptic transmission. The brown ovals were used to highlight the genes with p-value <0.001

bipolar disorder, and schizophrenia (Redies et al. 2012). An OCD GWAS found that cadherin 10, type 2 (CDH10) was reported as the second strongest association signal for OCD susceptibility (Mattheisen et al. 2014). Together, these findings suggest that cell-cell contact molecules may be involved in SRI response in OCD patients. Although, it should be mentioned that due to the lack of adequate biological evidence in OCD to support this, further study is needed.

A gene in the glutamatergic neurotransmission system, *GRIN2B*, an N-methyl-D-aspartate glutamate receptor, was one of the genes relevant to OCD and SRI response with other significant SNPs. Three previous genetic studies reported a significant association between a variant in *GRIN2B* and OCD (Arnold et al. 2004; Stewart et al. 2007; Alonso et al. 2012). Volumetric magnetic resonance imaging suggested that genetic variations of *GRIN2B* are associated with volumetric brain abnormalities in OCD (Arnold et al. 2009). It was also found that *GRIN2B* variations interact with the variations in *SLC1A1*, which is the susceptibility gene consistently replicated in OCD. Our GWAS analysis did not provide strong evidence for a single-variant association in the glutamatergic or serotonergic neurotransmission systems contributing to SRI response.

The enrichment analysis did indicate that multiple genes in the glutamatergic and serotonergic neurotransmission systems may jointly contribute to SRI treatment in OCD (Table 5.3 and Fig. S1). More of the nominated genes occur in the glutamatergic pathway than the serotonergic pathway (Fig. S1). We do recognize that our study is underpowered in detecting all neuropathic SNPs for enrichment.

Even though one genome-wide significance hit and two suggestive pathway enrichment scores were found, some potential imitations of this study should be mentioned. Firstly, drug response was based on a retrospective self-report. Second, since large OCD samples with drug information are rare, the analysis was based on a limited sample. Last, there was a lack of detailed information on both the dosage and duration of SRI medication, as well as the receipt of behavioral therapy. For future studies, it would be important to include measures of treatments in greater detail with response including reliable measures of symptom reduction within the first few months of treatment.

Several strengths of the study should be mentioned as well. First, the rigorous semi-structured clinical examination and diagnostic best-estimation procedures support phenotypic reliability. Second, given the clinical and genetic heterogeneity of OCD, the OCGAS sample attempted to increase homogeneity by targeting OCD-affected individuals with an early age of onset.

Up to 30% of OCD patients show minimal clinical improvement with treatment, which could indicate the biological heterogeneity of OCD phenotypes. Thus, it would be worthwhile to consider subgroups of OCD patients defined by their drug response. This could potentially provide a relatively more homogenous population (Davis et al. 2002). It is also worth noting that the study participants came from two studies, one of which was a family-based linkage study and the other was a trios-based association study. Relatedness may confound association tests and odds ratio estimation; the MQLS test offers a better way to conduct a test that corrects for the

relatedness coefficient with pedigrees, using a kinship matrix (identity by descent) from genotype data (Thornton et al. 2007).

Further research is needed to replicate the current findings on genetic variations related to SRI response in individuals affected by OCD. It is anticipated that next-generation sequencing methods, which facilitate the analysis of multiple genes, including the effects of both common and rate variants, will provide further understanding of OCD mechanisms of treatment response and lead to more effective forms of treatment for OCD (Korf and Rehm 2013).

References

Allen NJ, Bennett ML, Foo LC, Wang GX, Chakraborty C, Smith SJ, et al. Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via GluA1 AMPA receptors. Nature. 2012;486(7403):410–4. [PubMed: 22722203].

Alonso P, Gratacos M, Segalas C, Escaramis G, Real E, Bayes M, et al. Association between the NMDA glutamate receptor GRIN2B gene and obsessive-compulsive disorder. J Psychiatry Neurosci. 2012;37(4):273–81. [PubMed: 22433450].

American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-IV). Washington, DC: Psychiatric Press; 1994.

Arnold PD, Rosenberg DR, Mundo E, Tharmalingam S, Kennedy JL, Richter MA. Association of a glutamate (NMDA) subunit receptor gene (GRIN2B) with obsessive-compulsive disorder: a preliminary study. Psychopharmacology. 2004;174(4):530–8. [PubMed: 15083261].

Arnold PD, Macmaster FP, Hanna GL, Richter MA, Sicard T, Burroughs E, et al. Glutamate system genes associated with ventral prefrontal and thalamic volume in pediatric obsessive-compulsive disorder. Brain Imaging Behav. 2009;3(1):64–76. [PubMed: 21031159].

Brandl EJ, Muller DJ, Richter MA. Pharmacogenetics of obsessive-compulsive disorders. Pharmacogenomics. 2012;13(1):71–81. [PubMed: 22176623].

Brandl EJ, Tiwari AK, Zhou X, Deluce J, Kennedy JL, Muller DJ, et al. Influence of CYP2D6 and CYP2C19 gene variants on antidepressant response in obsessive-compulsive disorder. Pharmacogenomics J. 2014;14(2):176–81. [PubMed: 23545896].

Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25(5):655–61. [PubMed: 19098027].

Davis KL, Charney D, Coyle JT, Nemeroff C. Neuropsychopharmacology: the fifth generation of progress. Philadelphia: Lippincott Williams & Wilkins; 2002.

Di Bella D, Erzegovesi S, Cavallini MC, Bellodi L. Obsessive-Compulsive Disorder, 5-HTTLPR polymorphism and treatment response. Pharmacogenomics J. 2002;2(3):176–81. [PubMed: 12082589].

Etheridge LA, Crawford TQ, Zhang S, Roelink H. Evidence for a role of vertebrate Disp1 in long-range Shh signaling. Development. 2010;137(1):133–40. [PubMed: 20023168].

Ferguson JM. SSRI antidepressant medications: Adverse effects and tolerability. Prim Care Companion J Clin Psychiatry. 2001;3(1):22–7. [PubMed: 15014625].

Hindorff LAMJ, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. A catalog of published genome-wide association studies. 2013. URL: www.genome.gov/gwastudies

Jun KR, Hur YJ, Lee JN, Kim HR, Shin JH, Oh SH, et al. Clinical characterization of DISP1 haploinsufficiency: a case report. Eur J Med Genet. 2013;56:309–13.

Kim SY, Chung HS, Sun W, Kim H. Spatiotemporal expression pattern of non-clustered protocadherin family members in the developing rat brain. Neuroscience. 2007;147(4):996–1021. [PubMed: 17614211].

Korf BR, Rehm HL. New approaches to molecular diagnosis. JAMA. 2013;309(14):1511–21. [PubMed: 23571590].

Mattheisen M, Samuels JF, Wang Y, Greenberg BD, Fyer AJ, JT MC, et al. Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. Mol Psychiatry. 2014;20:337–44.

Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, Hill RS, et al. Identifying autism loci and genes by tracing recent shared ancestry. Science. 2008;321(5886):218–23. [PubMed: 18621663].

Murphy DL, Lesch KP. Targeting the murine serotonin transporter: insights into human neurobiology. Nat Rev Neurosci. 2008;9(2):85–96. [PubMed: 18209729].

Nestadt G, Grados M, Samuels JF. Genetics of obsessive-compulsive disorder. Psychiatr Clin North Am. 2010;33(1):141–58. [PubMed: 20159344].

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336–7. [PubMed: 20634204].

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. [PubMed: 17701901].

Qin H, Samuels JF, Wang Y, Zhu Y, Grados MA, Riddle MA, et al. Whole-genome association analysis of treatment response in obsessive-compulsive disorder. Mol Psychiatry Macmillan Publishers Limited. 2015;21:270. https://doi.org/10.1038/mp.2015.32.

Redies C, Hertel N, Hubner CA. Cadherins and neuropsychiatric disorders. Brain Res. 2012;1470:130–44. [PubMed: 22765916].

Samuels JF, Riddle MA, Greenberg BD, Fyer AJ, McCracken JT, Rauch SL, et al. The OCD collaborative genetics study: methods and sample description. Am J Med Genet B Neuropsychiatr Genet. 2006;141B(3):201–7. [PubMed: 16511842].

Sangkuhl K, Klein TE, Altman RB. Selective serotonin reuptake inhibitors pathway. Pharmacogenet Genomics. 2009;19(11):907–9. [PubMed: 19741567].

Shugart YY, Samuels J, Willour VL, Grados MA, Greenberg BD, Knowles JA, et al. Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q. Mol Psychiatry. 2006;11(8):763–70. [PubMed: 16755275].

Shugart YY, Wang Y, Samuels JF, Grados MA, Greenberg BD, Knowles JA, et al. A family-based association study of the glutamate transporter gene SLC1A1 in obsessive-compulsive disorder in 378 families. Am J Med Genet B Neuropsychiatr Genet. 2009;150B(6):886–92. [PubMed: 19152386].

Stewart SEFJ, Moorjani J, Jenike E, Beattie K, Illmann C, Delorme R, Leboyer M, Sedovic M, Smoller J, Jenike M, Pauls D. Family-based association between obsessive compulsive disorder and glutamate receptor candidate genes. New York: World Congress of Psychiatric Genetics; 2007.

Stewart SE, Yu D, Scharf JM, Neale BM, Fagerness JA, Mathews CA, et al. Genome-wide association study of obsessive-compulsive disorder. Mol Psychiatry. 2013;18(7):788–98. [PubMed: 22889921].

Tansey KE, Guipponi M, Perroud N, Bondolfi G, Domenici E, Evans D, et al. Genetic predictors of response to serotonergic and noradrenergic antidepressants in major depressive disorder: a genome-wide analysis of individual-level data and a meta-analysis. PLoS Med. 2012;9(10):e1001326. [PubMed: 23091423].

Taylor S. Molecular genetics of obsessive-compulsive disorder: a comprehensive meta-analysis of genetic association studies. Mol Psychiatry. 2013;18(7):799–805. [PubMed: 22665263].

Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasilikelihood score test. Am J Hum Genet. 2007;81(2):321–37. [PubMed: 17668381].

Voyiaziakis E, Evgrafov O, Li D, Yoon HJ, Tabares P, Samuels J, et al. Association of SLC6A4 variants with obsessive-compulsive disorder in a large multicenter US family study. Mol Psychiatry. 2012;16(1):108–20. [PubMed: 19806148].

Chapter 6
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

6.1 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, 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

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

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_6

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.

6.1.1 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 1100 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 6.1. 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 (V355 M) 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 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.

Table 6.1 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

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.

6.1.2 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.

6.1.3 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.

6.2 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. 6.1).

6.2.1 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

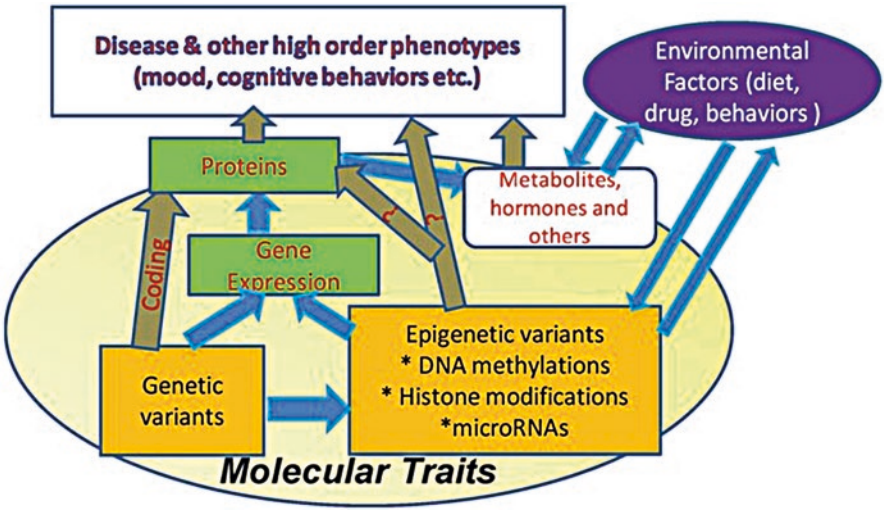


Fig. 6.1 Molecular traits link DNA/RNA to environment and high-order phenotypes

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. 6.2). 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), whole blood (Fehrmann et al. 2011), 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.

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 6.2.

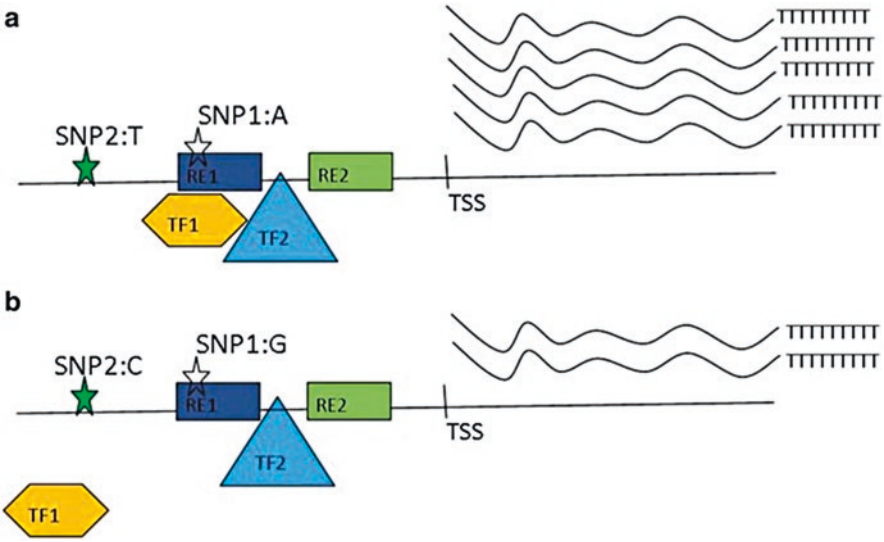


Fig. 6.2 Model of SNPs presenting eQTL in cis-association. SNP1 is located inside regulatory element1 (*RE1*). Its A allele has strong binding affinity to transcription factor1 (*TF1*). Its G allele does not bind TF1 well and consequently leading to reduced expression. SNP2 is in linkage disequilibrium with SNP1. Therefore, genotypes of both SNP1 and SNP2 show correlation with expression. *TSS* transcription start site

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.

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.

Table 6.2 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 500 K, 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 of Alzheimer patients (European descent)	Affymetrix 500 K, 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	2944–4781 cis-associations and 471–1826 trans-associations are significant (permutation correction for SNPs tested, FDR for traits tested)

(continued)

Table 6.2 (continued)

Authors	Samples	Platforms	Findings with definition of significance
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)
Colantuoni et al. (2011)	269 individuals, prefrontal cortex, (primary African American and Caucasian samples)	Either Illumina Infinium II 650 K or Illumina Infinium HD Gemini 1 M Duo BeadChips for genotype; custom expression array	1628 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	6794 significant cis-eQTLs, 991 significant trans-eQTLs in parietal cortex region and 9010 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

6.2.1.1 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 8000-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 1000 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 to be used more in eQTL mapping studies.

6.2.1.2 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).

6.2.1.3 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 carefully controlling the population structure issue, but it will overestimate power for detecting population-unique eQTLs.

6.2.1.4 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 require larger sample collection. With 1469 unrelated blood samples, high-quality trans-associations were detected and replicated in a different set of tissues and sample collections (Fehrmann et al. 2011).

6.2.1.5 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.

Both sample demographic information and clinical measures are 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.

6.2.1.6 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.

6.2.1.7 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) has 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 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.

6.2.2 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 6.3. The methylation level is quantified as percentage of methylation at a specific CpG site, with values ranging from 0 to 1.

Figure 6.3 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 3000 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 6.3 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 2878 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	7966–12,081 cis-mQTLs, 2893–4653 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; Illumina 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

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.

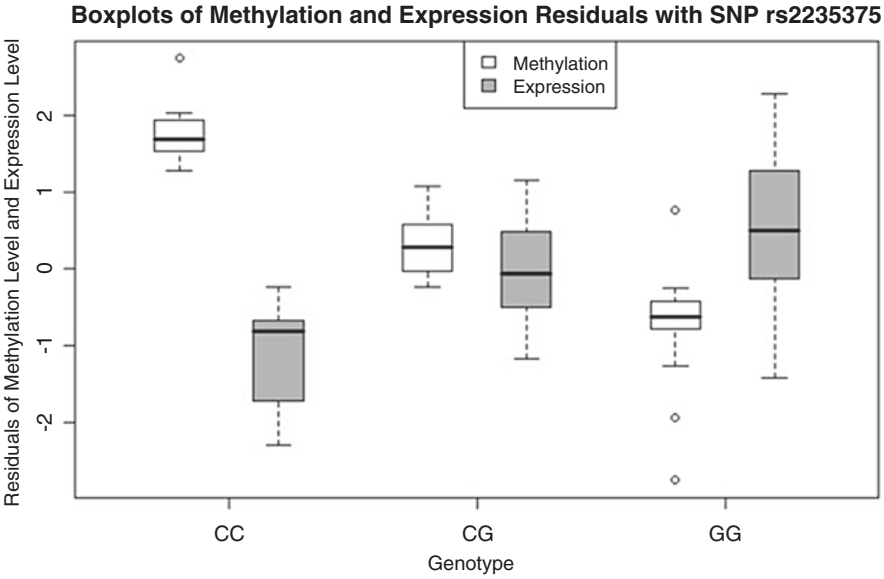


Fig. 6.3 DNA methylation and gene expression of IRF6 is correlated with genotypes of rs2235375. DNA methylation and gene expression are negatively correlated. (From Zhang et al. 2010)

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), and 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.

6.3 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 $p < 4e{-}8$ (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 ($p < 1e{-}10$) (Hicks et al. 2009).

6.4 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.

6.5 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.

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.

6.6 Database or 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.

Acknowledgment 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):1210–8.
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):411–9.

Chapter 7
From Family Study to Population Study: A History of Genetic Mapping for Nasopharyngeal Carcinoma (NPC)

Haide Qin and Yin Yao

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 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*

H. Qin
Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, People's Republic of China

Y. Yao (✉)
Unit of Statistical Genomics, Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA
e-mail: kay1yao@mail.nih.gov

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_7

7.1 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.

7.2 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 (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.

7.3 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.

7.3.1 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.

7.3.2 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 7.1). 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).

7.3.3 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 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

Table 7.1 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 (2002a)	*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;	NA
								1.58, 3.10	
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 (seven 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 (2005)	*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	Ben Nasr (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 (2010a, b)	*79*	156	161	VEGF	VEGF	1.65	1.05, 2.58	0.029
	Wei (2007a, b, c)	*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 (2007a, b)	*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

(continued)

Table 7.1 (continued)

	Study (first author and year)	Ref	Cases	Cont.	Genes studied	Significant gene associations	Odds ratio	95% CI	*P* value
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
	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 (2010)	*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 (2010a, b)	*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 interaction 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

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 = 1568/1297), only two SNPs that were in strong LD with each other ($r2 = 0.7$) 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 = 9770/10,799) 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.

7.3.4 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.

7.3.5 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 1000 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).

7.3.6 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.

7.3.7 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.

7.3.8 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, b, 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 7.1). 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.

7.4 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 1583 cases and 1897 controls (in the discovery stage). Despite the differences in sample sizes, both studies identified their most significant signal in the *HLA* region (6p21) (Fig. 7.1).

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 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 non-neuronal 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

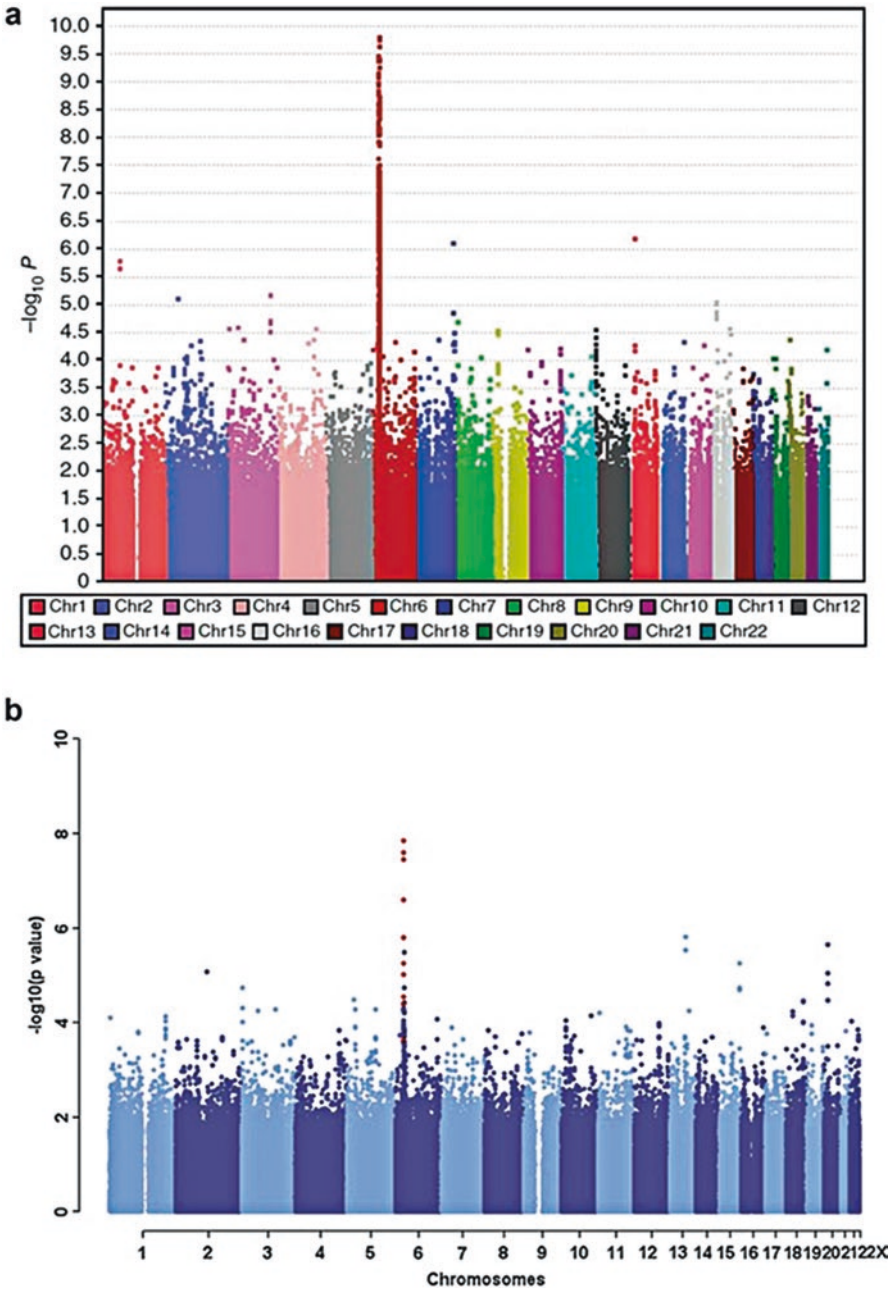


Fig. 7.1 GWAS showing evidence for the association of *HLA* and nasopharyngeal carcinoma risk. *Panel a*. Manhattan plot of the genome-wide *P* values of association for the mainland GWA study in Southern China (Bei et al. 2010). *Panel b*. Manhattan plot of the genome-wide *P* values of association for the GWA study in Taiwan (Tse et al. 2009)

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).

7.5 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 seems 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 of 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 con-

tributions 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 to 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.

Acknowledgments 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):114–9.

Chang ET, Adami HO. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomark 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 Biomark 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 Biomark 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 Biomark 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 Investig. 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 Biomark 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 Biomark 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 Biomark 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 DG Jr, 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. 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 Biomark 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):1911–6.
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. T1D/Base: GABBR1. 2011. 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):116–9.
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):112–7.
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 Biomark 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 Biomark 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 8
Efficient Test for Nonlinear Dependence of Two Continuous Variables

McKenzie Ritter, Yi Li, Yi Wang, Yin Yao, and Li Jin

Abstract A new method to test nonlinear dependence between two continuous variables (*X* and *Y*) is proposed. This is achieved by using continuous analysis of variance (CANOVA). The software is available at https://sourceforge.net/projects/canova. First, a neighborhood for each data point related to its *X* value was defined. Then, the variance of the *Y* value within the neighborhood was calculated. Last, permutations to evaluate the significance of the observed values within the neighborhood variance were conducted. To examine the strength of CANOVA compared to six other methods, extensive simulations were completed to examine the false-positive rates and statistical power. Both simulation and real datasets (kidney cancer RNA-seq data) were used. From these analyses, it was concluded that CANOVA is efficient as a method in testing nonlinear correlation and has several advantages for real data application.

8.1 Background

Any statistical relationship between two random variables or sets of data is known as dependence. In contrast, correlation is any broad class of statistical relationships that include dependence. Correlation is typically useful in indicating a predictive relationship. Several methods exist that measure this degree of correlation. For example, the Pearson correlation coefficient is the most commonly used method. It is only sensitive to linear correlations, but several other methods exist that are more

M. Ritter · Y. Yao
Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA

Y. Li · Y. Wang · L. Jin (✉)
Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China
e-mail: lijin@fudan.edu.cn

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_8

robust for the use of nonlinear correlations (Croxton and Cowden 1939; Dietrich 1991; Aitken 1942). The Pearson correlation coefficient ranges from −1 to 1, developed by Karl Pearson from a related idea of Francis Galton's (1877, 1886; Lockyer 1885; Pearson 1895; Stigler 1989). The Pearson correlation coefficient is the covariance of two variables divided by the product of their standard deviations. Even though Pearson's correlation is widely used, several negative effects are associated with it. These include a non-robust Pearson's r sample statistic (Horn 1998) and the potential for misleading values when outliers are present (Devlin et al. 1975; Huber 2011). The alternative hypothesis for the Pearson correlation test is the linear correlation between two variables X and Y.

The two most common nonlinear rank-based correlation coefficients are Spearman's rank correlation and Kendall's rank correlation. Spearman's rank correlation coefficient (or Spearman's rho) is a nonparametric measure of statistical dependence between two variables. It is defined as the Pearson correlation coefficient between the ranked variables (Myers et al. 2010). The Kendall rank correlation coefficient (or Kendall's tau) is used to test the coefficient between two measured variables (Kendall 1938). This test is nonparametric because it does not rely on any assumption of the distribution of X or Y. The alternative hypothesis for both Spearman's and the Kendall correlation test states that the correlation between variables X and Y corresponds to a monotonic function.

There are other commonly used methods that measure the correlation between random variables that include distance correlation, Hoeffding's independence test, Maximal information coefficient (MIC), Hilbert-Schmidt Information Criterion (HSCI), and Heller-Heller-Gorfine distance (HHG). The distance correlation is a measure of statistical dependence between two arbitrary variables or random vectors. It was developed by Gabor J. Szekely in 2005 to address the deficiency of Pearson's r (which can be equal to zero-dependent variables). The initial results for distance correlation were published in 2007 and 2009 (Székely et al. 2007; Kosorok 2009). The distance correlation is zero if the random variables are statically independent. If the distance correlation is one, the dimensions of the linear spaces spanned by X and Y are almost equal, and Y is a linear function of X. Hoeffding's independence test is based on the population measure of deviation from independence. A sample-based version of this measure was described with a calculation under the null distribution in 2008 (Wilding and Mudholkar 2008). If the continuous joint distribution and marginal probability densities of two random variables exist, the Hoeffding's independence test will be efficient. MIC is a measure of the degree of linear or nonlinear association between two random variables. This is a nonparametric method based on maximal information theory (Reshef et al. 2011). This method uses binning to apply mutual information to continuous random variables. Binning previously was used to apply mutual information to continuous distributions. MIC is a method for selecting the number of bins and finding a maximum over possible grids. Even though there are merits of MIC, there are several limitations of this method. Specifically, the approximation algorithms with better time-accuracy tradeoffs should be used (Reshef et al. 2013). The hypothesis of MIC has a wide range of associations. HSIC contains an independence criterion based on the eigenspec-

trum of covariance operators in reproducing kernel Hilbert spaces (RKHSs) that consist of an empirical estimate of the Hilbert-Schmidt Independence Criterion (Gretton et al. 2005). HHG is a powerful test that is applicable to all dimensions, consistent against all alternatives, and easy to implement (Heller et al. 2012).

We will focus on the alternative hypothesis, which says "similar X values lead to similar Y values," or more formally, $Y = f(x) + e$, $e \sim N(0, s)$, $s > 0$, where f is a non-constant smooth function. A novel nonlinear correlation measure method named continuous analysis of variance (CANOVA) test is proposed. The traditional analysis of variance (ANOVA) uses categorical factors, which were used as a base for CANOVA (Scheffe 1999). ANOVA tests whether the variance either within or between categories is smaller or greater than what is randomly expected. When encountering continuous response with continuous factors, a generalization of the "within category variance" is needed for ANOVA. Thus, with the use of CANOVA, a neighborhood of each data point according to its X value is defined, and then the variance of Y is calculated within that neighborhood. A permutation test is then conducted to determine the significance of the observed "within neighborhood variance." CANOVA was compared to six other methods using simulated data. The false-positive rate (Burke et al. 1988) and statistical power (Cohen 1988) for CANOVA, as well as the other size methods, were checked using both simulated and real datasets (RNA-seq data on kidney cancer) (Jiang et al. 2014; The Cancer Genome Atlas Research Network 2013). This chapter is based on a previously published manuscript (Wang et al. 2015).

8.2 Methods

The two random variables X and Y are denoted X_i and Y_i for the ith observation. The within neighborhood sum square statistics are defined as:

$$W = \sum_{i,j} \left(Y_i - Y_j\right)^2, j < i, \left|rank\left(X_i\right) - rank\left(X_j\right)\right| < K \tag{8.1}$$

where K is an integer constant provided by the user. It should be noted that $|\text{rank}(X_i) - \text{rank}(X_j)| < K$ defines the neighborhood of the dataset. Again, the alternative hypothesis for CANOVA states that "similar or neighboring X values will lead to similar Y values." So, when X and Y are correlated, the W statistics tend to be smaller than what would randomly be expected. To evaluate the significance of the observed W, a permutation test is performed (Good 2000). When the X values are equal (tie), the rank of the tied X values are randomly shuffled with each permutation. For example, if the data were $X = 1, 1, 2, 3$ and $Y = 2, 1, 7, 4$, X has two ones, so the sorting of the data points is not unique. The algorithm randomly chooses a sorting pattern, which could be $X = 1, 1, 2, 3$; $Y = 2, 1, 7, 4$; or $X = 1, 1, 2, 3$; $Y = 1, 2, 7, 4$. The algorithm is implemented by the CANOVA software in the Linux

system (available at httpys://sourceforge.net/projects/canova/). The CANOVA algorithm (pseudocode) is summarized as:

```
sort data points according to X value

for (i = 0; i < #tie_shuffe; i + +) {

shuffle Y of tied X values

calculate observed W_i using observed Y}

Observe W = average (W_i)

Count = 0;

for(i = 0; < #permutations; i + +)

{ calculate random W using random shuffled Y

if (random W ≤ observed W) count + +}

return p – value = count/#permutations
```

When calculating W, we take advantage of the fact that X_i is sorted. Thus, the algorithm complexity is $O(n\log n + np)$, where n is the sample size and p is the number of permutations. When testing many Xs against one Y, only one permutation of Y is needed that can be reused for all Xs.

8.3 Simulation Study

Nine simple functions were simulated and added the Gaussian noise (mean = 0, variance = 1) to the Y value of each function (Table 8.1). This included constant functions (i.e., a linear function of the form $y = b$, where b is a constant and $b = 0$ in Table 8.1), linear functions, quadratic functions, sine functions, and cosine functions. The Gaussian noise levels were varied (mean = 0; variance = 1/9, ¼, 4, and 9) for the simulations. The power was then reported across the noise levels (which can be found in additional file 1 from the original manuscript) (Wang et al. 2015). The six other methods were benchmarked, which included Pearson correlation coefficient, Spearman's rank correlation coefficient, Kendall's rank correlation coefficient, distance correlation, Hoeffding's independence test, and the maximal information coefficient. The simulation was repeated 1000 times to calculate the false-positive rate and statistical power. Fifty were chosen as the sample size ($N = 50$), with x as the independent variable, which was uniformly distributed

Table 8.1 Simulation power in nine simple functions

$N = 50$, x~U(−1.1)	CANOVA2	CANOVA4	CANOVA8	CANOVA12	Pearson	Kendall	Spearman	Distance	Hoeffding	MIC
$y = 0 + N(0,1)$	0.051	0.048	0.048	0.050	0.047	0.048	0.049	0.039	0.059	0.051
$y = 0 + N(0,1)$	0.564	0.798	0.889	0.902	**0.972**	0.962	0.961	0.950	0.953	0.591
$y = 0.5;^{*}(x + 1)^2 + N(0,1)$	0.606	0.836	0.904	0.918	**0.968**	0.953	0.962	0.964	0.953	0.633
$y = \sin(Pi^{*}x) + N(0,1)$	0.758	0.941	0.966	0.962	0.936	0.918	0.930	**0.969**	**0.969**	0.829
$y = \sin(2^{*}Pi^{*}x) + N(0,1)$	0.713	**0.886**	0.812	0.294	0.318	0.328	0.320	0.341	0.405	0.579
$y = \sin(3^{*}Pi^{*}x) + N(0,1)$	0.677	**0.796**	0.254	0.076	0.178	0.192	0.199	0.186	0.219	0.423
$y = \cos(Pi^{*}x) + N(0,1)$	0.784	0.940	**0.973**	0.942	0.067	0.076	0.083	0.660	0.710	0.660
$y = \cos(2^{*}Pi^{*}x) + \mathrm{N}(0,1)$	0.738	**0.891**	0.754	0.142	0.045	0.054	0.053	0.100	0.129	0.548
$y = \cos(3^{*}Pi^{*}x) + N(0,1)$	0.673	**0.751**	0.160	0.031	0.053	0.054	0.057	0.074	0.090	0.371

The bold means the first place result of all methods compared

Table 8.2 Power comparison in kidney cancer dataset (the significance level $\alpha = 0.05/20531$)

Kidney cancer dataset	CANOVA	Kendall	Pearson	Spearman	Hoeffding	Distance	MIC
Significant gene number	5901	11,569	8239	**11,629**	4953	10,946	8081
Competing time (seconds)	**24**	65	32	32	44	~10^6	114

The bold means the first place result of all methods compared
~ means about or approximately

in (−1, 1) and y as the dependent variable. K is the only parameter of CANOVA, and its value was assigned from the positive integer collection ($K = 2, 4, 8, 12$). To note, MIC also has a similar bias/variance parameter ("alpha" parameter in the Minerva implementation), which is the maximum allowed resolution of any grid (Reshef et al. 2011). Reshef et al. also found that the different parameter setting ($\alpha = 0.55$, $c = 5$) is faster than default and does not appear to significantly affect the performance (2013). To simplify things, the default parameters of MIC were used ($\alpha = 0.6$, $c = 15$).

8.4 Applications on Real Data

The proposed CANOVA method was applied to an RNA-seq kidney cancer dataset and was then compared to the results generated by the other six methods. The kidney cancer data used contained 604 samples, which included 20,531 genes (Jiang et al. 2014; The Cancer Genome Atlas Research Network 2013). The correlation between genotype data X (20,531 gene expression data) and phenotype data Y (kidney cancer or not) were tested. The significance after Bonferroni correction was set to 2.435342e-06. An X–Y plot and grid search (like, $K = (10, 20, 30, 40, 50)$) were used to choose the best K, which was $K = 30$ for CANOVA based on statistical power. The other methods used their default parameters (i.e., MIC, $\alpha = 0.6$, $c = 15$). Table 8.2 lists the results, as well as the comparisons of these methods.

8.5 Results

8.5.1 Results from the Simulation Study

As shown in Table 8.1, when the constant function of $y = 0$ was used, the false-positive rate of the six different methods using alpha = 0.05 as the significance level was used. CANOVA, using different K values (CANOVA2, CANOVA4, CANOVA8, CANOVA12), Pearson's correlation coefficient, Spearman's rank correlation coefficient, Kendall's rank correlation coefficient, and the Maximal information coefficient, all had false-positive rates around 0.05, which indicates correct results. The

distance correlation's false-positive rate fell slightly below 0.05, and Hoeffding's independence test's false-positive rate was a little above 0.05. It is important to note that the significant variables in Hoeffding's independence test have the potential to be false positives and the actual significant variables may not have been detected by the distance correlation.

In terms of power for the nonconstant correlations (Table 8.1), the following were seen: (1) when the correlation is linear, the Pearson correlation coefficient is the most powerful. CANOVA is less powerful than Pearson but does not fail (power > 0.5). (2) With nonlinear correlation, CANOVA is the best, especially when the correlation is highly oscillating/nonlinear. (3) The power of CANOVA4 is the best single nonlinear test and more powerful than MIC with sine and cosine functions.

The power comparison for nonconstant correlations yielded results that can be found in additional file 1 from the original manuscript (Wang et al. 2015). The results are as follows: (1) when the Gaussian noise levels were low (Gaussian variance = 1/9, ¼), most of the methods had higher power, specifically with simple linear relationships. CANOVA2 and CANOVA4 were two of the better methods with high power in most of the nonconstant functions. (2) When the Gaussian noise levels were high (Gaussian variance = 4, 9), most of the methods had low power, while CANOVA4 had higher power in simple linear relationship functions. Thus, when the correlation between the two random variables is linear, Pearson correlation coefficient would be best to use to increase the statistical power. When there is a nonlinear or complicated correlation, CANOVA with parameter K is a good method to explore the data's correlation structure.

8.6 Results from the Kidney Cancer Study

Table 8.2 compares power and computing time for the kidney cancer data (Jiang et al. 2014; The Cancer Genome Atlas Research Network 2013). In order to compute time comparison, the number of permutations of CANOVA was set at 10,000,000 (Table 8.2). Table 8.3 shows the genes that were detected only by the CANOVA methods (not detected by the other methods). The genes only detected by other methods are listed in additional file 2, which can be found in the original manuscript (Wang et al. 2015). To explore the relationships found through CANOVA, the scatterplot and probability density distributions for gene expressions between the cases and controls were examined (Fig. 8.1). All of the CANOVA results were completed in the C++ environment (Stroustrup 1995), and the six other methods were calculated using the *R* package "energy" (Székely and Rizzo 2013), "Hmisc" (Hmisc, Harrell Miscellaneous), and "Minerva" (Albanese et al. 2013). The CANOVA results were parallelly (fully using all 8 CPU cores) calculated using a PC with an AMD FX-8320 CPU and 32GB memory. In addition, all of the *R* code was parallelly computed using the package "snow" (Tierney et al. 2009).

Table 8.3 Significant genes detected only by CANOVA and corresponding *p*-value of all methods in kidney cancer data ($\alpha = 0.05/20531$)

CANOVA_gene	CANOVA	Distance	Hoeffding	Kendall	Pearson	Spearman	MIC
ACY3|91703	0	4.00E-06	0.47918	0.286872598	0.002263414	0.287245869	0.189931316
AMD1|262	0	4.40E-05	0.08116	0.005927801	0.733642545	0.005833208	0.212586042
AMDHD1|144193	3.40E-07	8.00E-06	0.67325	0.030092326	0.000717698	0.029975253	0.170029851
C17orf37|84299	5.80E-07	5.20E-05	0.04005	3.6 IE-05	0.417383349	3.24E-05	0.219216883
C21orf57|54059	2.40E-07	4.00E-06	0.04784	6.30E-05	3.99E-05	5.38E-06	0.19141914
CRAT|1384	5.80E-07	8.00E-06	0.32615	0.000160458	3.77E-06	0.000149343	0.196028813
ETV5|2119	0	0.000172	0.42256	0.001755105	0.003401714	0.001702658	0.202086913
FAH|2184	0	0.000933998	0.48933	0.153797268	0.457070256	0.153962124	0.212691814
FAM105A|54491	0	2.00E-05	0.72088	0.005901803	768E-05	0.005807373	0.198623556
FTL|2512	0	0.002467995	0.4743	0.048315704	023060211	0.048231442	0.212746271
GDA|9615	1.60E-07	0.00025	0.48634	0.160122916	0.459724584	0.160300937	0.185681164
HSD17B14|51171	0	8.20E-05	0.19284	0.001051631	0.006799576	0.001012728	0.208298029
LOC100132111|100132111	1.00E-08	1.40E-05	0.08222	0.001103681	0.357830837	0.001063627	0.20892751
MCM3|4172	0	6.00E-06	0.50714	0.033769054	1.98E-05	0.033657199	0.197222887
MSL3L2|151507	5.00E-08	4.00E-06	0.0658	0.00022671	0.000309573	0.000212513	0.197191821
NPEPPS|9520	6.30E-07	1.80E-05	0.12107	0.006358039	0.294981611	0.006260864	0.193740442
RASEF|158158	4.00E-08	2.00E-05	0.15806	0.038695575	0.339964949	0.038592039	0.221013132
RASGRF1|5923	4.5C6 07	0.000509999	0.29384	0.005697491	0.944454242	0.005604368	0.192676281
SLC9A3R1|9368	0	6.00E-06	0.2351	0.001772375	0.000600274	0.001719639	0.211044758
SRGAP2|23380	1.49E-06	1.60E-05	0.13479	0.00010228	0.00076986	9.43E-05	0.16085031
SYTL2|54843	9.40E-07	0.00357999	0.49524	0.156725188	0.013347293	0.156896177	0.197737514
UGT1A9|54600	5.00E-08	1.60E-05	0.5995	0.278490278	133E-05	0.278854528	0.18041022
ZNF280B|140883	680E-07	0.000431999	0.17067	0.073453284	0.259146202	0.073428429	0.203602346
ZNF577|84765	0	4.60E-05	0.13197	0.063783754	0.410213566	0.063735193	0.208902832

As the *p*-value of MIC is calculated by table lookup, we just list the MIC value (if MIC >0.22378, then the *p*-value of MIC <2.435342e-06). The genes reported in PubMed were shown in bold italics

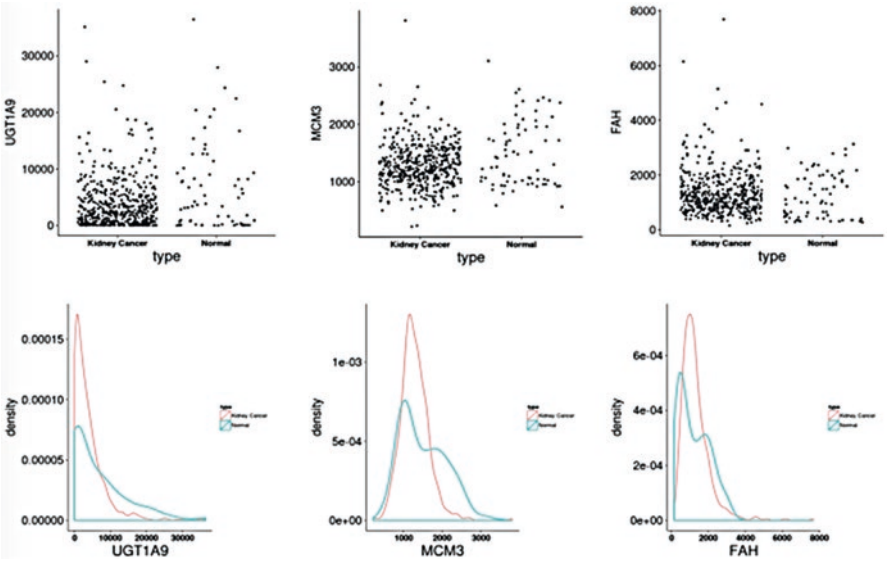


Fig. 8.1 The scatterplot and probability density distribution of three gene expressions (FAH, MCM3, and UGT1A9) between kidney cancer and normal groups

By using the kidney cancer RNA-seq data, Table 8.2 showed that Spearman detected the greatest number of significant genes ($\alpha = 0.05/20,531$) and CANOVA was the fastest method while using the desktop PC. In order to further explore the biological relevance of the genes that were detected and to compare the features of each of the methods, the uniquely "significant" genes from each method were used as the target gene set. A literature review for validation of each gene was then completed. The uniquely significant genes detected by CANOVA, as well as the corresponding p-values of each of the methods, can be seen in Table 8.3. The genes reported in PubMed (indicating the presence of abstracts discussing a relationship between kidney cancer and the gene) are listed in bold. The uniquely significant genes from the other methods can be found in additional file 2 in the original manuscript (Wang et al. 2015).

Of the unique set of genes detected by CANOVA, a few were of significance to kidney cancer/disease (Table 8.3): FAH, MCM3, and UGT1A9. Specifically, a defect in FAH results in the accumulation of FAA, which can lead to both oxidative stress and severe liver and kidney disease (Li et al. 2012; Dieter et al. 2003). The MCM3 gene was overexpressed in various cancers, which included kidney cancer (Ha et al. 2004). The UGT1A9 gene was found to be a major contributor to glucoronidation in the liver and kidney (Grosse et al. 2013).

MCM3 and FAH in Fig. 8.1 show that if the normal group distribution is bimodal and the level of expression is mild, then an individual is more likely to have kidney cancer. For FAH specifically, the mean kidney cancer distribution approaches normal, indicating the linear relationship is close to zero (Pearson R's p-value is

approximately 0.5 in Table 8.3). If the distribution is not bimodal, CANOVA can provide sufficient power if the two distributions have the same mean, but differing variances. For example, if the control group has a wider distribution (lower peaks), it will have a thicker tail on the left and right sides. Thus, higher or lower expression induces protection from the disease, like in UGT1A9 (Fig. 8.1).

The only unique gene detected by the distance method (reported in PubMed as well) is 1IGF1R. IGF1R was found to be indirectly associated with kidney cancer tumor growth (Zhang et al. 2013). A single gene was detected by the MIC method (reported in PubMed too), GIPC2. GIPC2 was reported to be downregulated in both kidney and colorectal tumors (Krikoshi and Katoh 2002). The only genes unique to the Pearson method (reported in PubMed) were EGR2 and COMT. EGR2's upregulation results in an overexpression of embryonic kidney cells in humans, which are indirectly associated with Wilms' tumors (Natrajan et al. 2006). The COMT polymorphism was found to be associated with renal cell cancer (Tanaka et al. 2007). It also should be noted that neither the Hoeffding or Kendall methods detected any unique genes.

8.7 Discussion

The proposed CANOVA method can be seen as an extension of ANOVA for continuous variables. The neighborhood is defined first and the within neighborhood variance is calculated. This is analogous to ANOVA's within treatment variance. The alternative hypothesis of CANOVA is that "similar X values lead to similar Y values." By calculating the variance of the Y values of similar or neighbor X values, the proposed hypothesis can be tested against the null.

Local regression is a method that is closely related to CANOVA because both estimate the local residual (Clevland et al. 1988). The statistical power would be expected to be similar. Specifically, if we took a moving average of every K point and then computed the R^2 between the estimated regression function and data, with this condition, two issues must be considered: (1) when K is an even number, a special treatment of the regression expectation on each data point is needed. (2) For the boundaries' data points, some special treatment is needed in order to calculate the unbiased regression expectation. The K nearest neighbor (kNN) regression is another type of local regression that is analogous to CANOVA (Altman 1992). CANOVA uses K as the parameter that defines the nearest neighbor of each of the data points. CANOVA tests the fitness of the neighborhood model, similar to kNN. Because Pearson's correlation coefficient can be seen as the model fitness test for a linear regression model, and CANOVA can be seen as analogous to the model fitness test of the kNN model, when using CANOVA, the permutation of one Y variable can be conducted, and then the association tests against numerous (i.e., 20,000) X variables can be completed with limited time. For kNN, the generated neighborhood for each X variable is different, so we must perform a permutation test for every combination of X and Y, potentially making kNN a slower method than

CANOVA. Also, CANOVA has the advantage of using direct independence testing, rather than having to use the regression step. Since the regression function does not need to be estimated, CANOVA is simpler and more elegant and thus preferred over local regression.

The W statistics distribution is unknown. For the simplest case, where $K = 2$, $Y \sim N(0, 1)$, and $W_2 = \sum_{i>1} (Y_i - Y_{i-1})^2$, we know that the mean $(W_2) = 2N - 2$ and var. $(W_2) = 12N - 16$ (calculated by Maple), where N is the sample size. In recognition of this, W does not follow a familiar distribution. A permutation test was used to assess the significance level. It only takes several seconds for a few hundred samples with 10^6 permutations on a desktop PC, with AMD FX-8320 CPU and 32 GB memory. CANOVA is even faster than Pearson's when testing the correlation between thousands of features and the one response variable Y. This elevated speed is because (1) CANOVA is implemented with C++ code, while Pearson's uses relatively slow R and (2) CANOVA is paralleled and uses all CPU cores, resulting in an 8X speed up on the AMD 8 core CPUs. (3) When testing 20,000 X variables against one Y variable, only one permutation test on the Y is needed, and then the permutation results for all X variables can be reused. The computational complexity is $O(np + \#Xn\log(n))$, where p is the number of permutations, $\#X$ is the number of X variables, and n is the sample size. This allows the framework to have potential applications for big data.

A parameter K is needed for CANOVA before beginning the test. It is up to the discretion of the user to pick a reasonable K. A large K gives more power for slow-varying functions, while a smaller K has more power for quick-oscillating functions. Thus, it is important that the user has prior knowledge of the function that is being tested. An X–Y plot would be useful to examine the data before testing. It is suggested that K = sample size/20 be used as a guide. The significance level must be preset (0.05/feature numbers), and then a grid search is used (such as K = (2, 30, 40, 80, 100, 200)) to choose the best K based on corresponding statistical power. Another method could be used, such as Pearson's or MIC, so one could get a better idea of the data. This information would allow one to choose a reasonable K for CANOVA.

CANOVA and MIC can be used to test nonlinear correlation, but CANOVA has specific advantages. While MIC tests all types of nonrandom correlations, CANOVA tests the alternative hypothesis, which says "similar X values lead to similar Y values." CANOVA's hypothesis is $Y = f(X) + e$, $e \sim N(0,s)$, $s > 0$, and f is a nonconstant smooth function. If the X and Y relationship cannot be written as $Y = f(X)$, then CANOVA could potentially fail. For example, for the relationship, $X^2 + Y^2 = 1$, CANOVA will fail, but MIC will work. CANOVA serves a purpose of offering a test of independence. The maximal information coefficient is used primarily as a measure of effect size and provides similar scores for relationships of similar strength, regardless of the relationship type (Reshef et al. 2011). Measurements of effect size can be used to test for independence (using a null hypothesis of zero effect size), but the reverse of this statement is not true. Justin B. Kinney and Gurinder S. Atwal indicated that the MIC method does not have the property of "equitability" and the simulation results contain artifacts (2014). Although, Reshef et al. (2014) and

Murrell et al. (2014) have speculated about the Kinney and Atwal's methodology. This work led to an overall better understanding of equitability and MIC, allowing researchers in this field to move forward.

The CANOVA method is less powerful than Pearson's in terms of linear correlation. This is a tradeoff between the hypothesis space and the statistical power. Pearson's has a very specific alternative hypothesis space (linear correlation), while the alternative hypothesis for CANOVA is more general. Many correlations are linear or approximately linear, making Pearson, Spearman, and Kendall correlation quite powerful.

The results of the kidney cancer analysis are shown in Table 8.3. Even though CANOVA was not able to detect the largest number of unique significant genes, it did find the largest number (three) of genes that were also found to be relevant to kidney cancer which was previously found in the literature.

The results of these three gene expressions (FAH, MCM3, and UGT1A9) showed that CANOVA could detect the special nonlinear relationships (Fig. 8.1, Table 8.3), which other methods could not find. The three genes were also reported to be biologically relevant in kidney cancer development (Li et al. 2012; Dieter et al. 2003; Ha et al. 2004; Grosse et al. 2013; Zhang et al. 2013).

It is known that each method has their own advantages, so the results of the different methods are often correlated. The results of the simulation indicated that using linear correlation (Pearson, Spearman, or Kendall), as well as a nonlinear correlation (CANOVA, MIC, Hoeffding, or Distance), could increase the odds of detecting biologically significant signals. In conclusion, based on our analyses, CANOVA seems to be efficient at testing nonlinear correlations and is applicable with real data.

8.8 Availability of Supporting Data

The kidney RNA-seq data was downloaded from the TCGA datasets (level 3 in the TCGA datasets: http://cancer-genome.nih.gov/).

References

Aitken AC. Statistical mathematics. Edinburgh: Oliver and Boyd; 1942.

Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C. Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics. 2013;29(3):407–8.

Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.

Burke DS, Brundage JF, Redfield RR, Damato JJ, Schable CA, Putman P, et al. Measurement of the false positive rate in a screening program for human immunodeficiency virus infections. N Engl J Med. 1988;319(15):961–4.

Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting–methods, properties, and computational algorithms. J Econ. 1988;37(1):87–114.
Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale: L Erlbaum Associates; 1988.
Croxton FE, Cowden DJ. Applied general statistics. New Jersey: Prentice-Hall Inc.; 1939.
Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation and outlier detection with correlation-coefficients. Biometrika. 1975;62(3):531–45.
Dieter MZ, Freshwater SL, Miller ML, Shertzer HG, Dalton TP, Nebert DW. Pharmacological rescue of the 14CoS/14CoS mouse: hepatocyte apoptosis is likely caused by endogenous oxidative stress. Free Radic Biol Med. 2003;35(4):351–67.
Dietrich CF. Uncertainty, calibration and probability: the statistics of scientific and industrial measurement. Boca Raton: CRC Press; 1991.
Galton F. Typical laws of heredity. 1877. 5.
Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst Great Brit Ireland. 1886;15:246–63.
Good P. Permutation tests. New York: Springer; 2000.
Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic learning theory. Heidelberg: Springer; 2005. p. 63–77.
Grosse L, Campeau AS, Caron S, Morin FA, Meunier K, Trottier J, et al. Enantiomer selective glucuronidation of the non-steroidal pure anti-androgen bicalutamide by human liver and kidney: role of the human UDP-glucuronosyltransferase (UGT)1A9 enzyme. Basic Clin Pharmacol Toxicol. 2013;113(2):92–102.
Ha SA, Shin SM, Namkoong H, Lee HJ, Cho GW, Hur SY, et al. Cancer-associated expression of minichromosome maintenance 3 gene in several human cancers and its involvement in tumorigenesis. Clin Cancer Res. 2004;10(24):8386–95.
Heller R, Heller Y, Gorfine M. A consistent multivariate test of association based on ranks of distances. Biometrika 2012:ass070.
Hmisc: Harrell Miscellaneous. http://CRAN.R-project.org/package=Hmisc.
Horn PS. Introduction to robust estimation and hypothesis testing. Technometrics. 1998;40(1):77–8.
Huber P. Robust statistics. In: Lovric M, editor. International encyclopedia of statistical science. Berlin/Heidelberg: Springer; 2011. p. 1248–51.
Jiang J, Lin N, Guo S, Chen J, Xiong M. Methods for joint imaging and RNA-seq data analysis. arXiv preprint. 2014;arXiv:14093899.
Kendall MG. A new measure of rank correlation. Biometrika. 1938;30:81–93.
Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci U S A. 2014;111(9):3354–9.
Kirikoshi H, Katoh M. Molecular cloning and characterization of human GIPC2, a novel gene homologous to human GIPC1 and Xenopus Kermit. Int J Oncol. 2002;20(3):571–6.
Kosorok MR. On Brownian distance covariance and high dimensional data. Ann Appl Stat. 2009;3(4):1266–9.
Li B, Reed JC, Kim HR, HJ C. Proteomic profiling of differentially expressed proteins from Bax inhibitor-1 knockout and wild type mice. Mol Cells. 2012;34(1):15–23.
Lockyer N. Nature: Macmillan Journals Limited. 1885.
Murrell B, Murrell D, Murrell H. R2-equitability is satisfiable. Proc Natl Acad Sci. 2014;111(21):E2160.
Myers JL, Well AD, Lorch RF Jr. Research design and statistical analysis. New York: Routledge; 2010.
Natrajan R, Little SE, Reis-Filho JS, Hing L, Messahel B, Grundy PE, et al. Amplification and overexpression of CACNA1E correlates with relapse in favorable histology Wilms' tumors. Clin Cancer Res. 2006;12(24):7284–93.
Pearson K. Note on regression and inheritance in the case of two parents. Proc R Soc Lond. 1895;58(347–352):240–2.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
Reshef D, Reshef Y, Mitzenmacher M, Sabeti P. Equitability analysis of the maximal information coefficient, with comparisons. arXiv preprint. 2013;arXiv:13016314.
Reshef DN, Reshef YA, Mitzenmacher M, Sabeti PC. Cleaning up the record on the maximal information coefficient and equitability. Proc Natl Acad Sci. 2014;111(33):E3362–3.
Scheffe H. The analysis of variance, vol. 72. New York: Wiley; 1999.
Stigler SM. Francis Galton's account of the invention of correlation. Stat Sci. 1989;4:73–9.
Stroustrup B. The C++ programming language: Pearson Education India. 1995.
Székely GJ, Rizzo ML. Energy statistics: a class of statistics based on distances. J Stat Plan Inference. 2013;143(8):1249–72.
Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35(6):2769–94.
Tanaka Y, Hirata H, Chen Z, Kikuno N, Kawamoto K, Majid S, et al. Polymorphisms of catechol-O-methyltransferase in men with renal cell cancer. Cancer Epidemiol Biomark Prev. 2007;16(1):92–7.
The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43–9.
Tierney L, Rossini A, Li N. Snow: a parallel computing framework for the R system. Int J Parallel Prog. 2009;37(1):78–90.
Wang Y, Li Y, Cao H, Xiong M, Shugart YY, Jin L. Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics. 2015;16(1). https://doi.org/10.1186/s12859-015-0697-7.
Wilding GE, Mudholkar GS. Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions. Stat Meth. 2008;5(2):160–70.
Zhang T, Niu X, Liao L, Cho EA, Yang H. The contributions of HIF-target genes to tumor growth in RCC. PLoS One. 2013;8(11):e80544.

Chapter 9
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.

Keywords Complex disease · Exome sequencing · Mendelian disease · Sequence alignment · Variant annotation

A. Collins (✉)
Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, University of Southampton, Southampton, UK
e-mail: a.r.collins@soton.ac.uk

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_9

9.1 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.

9.2 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.

9.2.1 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.

9.2.2 De Novo Variants

F or '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.

9.2.3 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.

9.2.4 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 (2000 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.

9.2.5 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.

9.3 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 $Q\text{sanger} = -10\log 10p$. 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.

9.3.1 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.

9.3.2 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.

9.3.3 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 (1000 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.

9.3.4 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 non-causal 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 $X\mathrm{j} = 1$ 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.

9.3.5 Copy Number Variant (CNV) and Loss of Heterozygosity Analysis

Test for structural variation has been typically undertaken using array comparative genome hybridisation (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).

9.3.6 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 ongoing and developing NGS 'data deluge'.

9.4 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:S12–7.
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, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011;48:580–9. https://doi.org/10.1136/jmedgenet-2011-100223.
McClellan J, King MC. Genetic heterogeneity and human disease. Cell. 2010;141:210–7.
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;27:2790–6. https://doi.org/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;27:2648–54. https://doi.org/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 10
Machine Learning Approaches: Data Integration for Disease Prediction and Prognosis

Andrew Collins and Yin Yao

Abstract Machine learning (ML) is an analytical approach that has been on increasing importance in this field. In this chapter, we would like to highlight the use of ML for disease risk prediction and prognosis to identify the scope of successful applications to date. Despite the enthusiasm, we feel that the evaluation of ML methods in real data sets has been limited thus 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 the era of next-generation sequencing.

10.1 Applications of Machine Learning

Enormous volumes of genomic data encompassing diverse data types (including gene expression, genetic polymorphisms, structural mutations, DNA methylation, eQTLs, and proteomic data) can 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) that resulted in 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

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

Y. Yao (✉)
Unit of Statistical Genomics, Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA
e-mail: kay1yao@mail.nih.gov

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_10

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 tumors (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 bioinformatics 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, modeling 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 weaknesses, 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 nonlinear systems where there is a degree of variable interdependence, ML methods are subject to important limitations. Careful modeling 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. Nonlinear 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 increasingly being applied in disease prediction and prognosis models. Some recent applications, focusing on refining clinical counseling and treatment pathways, integrate epidemiological data and biomarker expression profiles. For example, Yu et al. (2010) developed 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 signaling 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. (2015) developed survival classifiers for NPC cases based on expression profiles of 18 tumor-associated biomarkers. The powerful classifier is focused on facilitating counseling and individualized patient management. Schulte et al. (2010) used SVM to predict survival for neuroblastomas 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 modeling 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 of 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) argued that GWAS analyses have ignored the full complexity of disease pathobiology, and the linear modeling 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 because 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).

References

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

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 Informat. 2006;2:59–77.

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

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.

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

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

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

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 Bioinform. 2012;14:251. https://doi.org/10.1093/bib/bbs024.

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 Y, Li Y, Cao H, Xiong M, Shugart YY, Jin L. Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics. 2015;16(1):260. https://doi.org/10.1186/s12859-015-0697-7.

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 11
OCD Genomics and Future Looks

McKenzie Ritter and Yin Yao

Abstract Obsessive compulsive disorder (OCD) has been studied using various genetic analyses over the years. It has been of interest to identify potential risk genes that point to the susceptibility of the disorder. Segregation analyses were the initial methods to study these genes. Linkage analyses were then used and slowly replaced segregation analyses in the genomics field. Now, genome-wide association studies (GWAS) and meta-analyses are commonly used to study OCD, as well as other psychiatric disorders. All previous research on OCD has focused on common variants, and the hope is to shift toward studying rare variants in the future. This chapter discusses each of these methodologies in the context of OCD, as well as a look into what the future of OCD statistical analyses may hold.

11.1 Introduction

Obsessive compulsive disorder (OCD) is a neurodevelopmental disorder that most commonly onsets in childhood. OCD is characterized by obsessions, which are persistent and unwanted thoughts, as well as compulsions, or repetitive behaviors that are done in response to the obsessions (Bokor and Anderson 2014). It is well documented that OCD is a highly heritable disorder, so it has been of interest in the field of statistical genomics (Davis et al. 2013; Katerberg et al. 2010).

The goal of these genetic analyses is to better understand the risk factors for the susceptibility of the disorder. Segregation analyses were used to study OCD, as well as other disorders. The analysis is not able to isolate specific regions or genes that act as risk factors for the inheritance of the disorder. Segregation instead seeks to identify the pattern of inheritance for a specific phenotype of the disorder (Elston 1981). This analysis requires the use of family structured data. There then was a shift in the field to linkage analyses.

M. Ritter · Y. Yao (✉)
Unit of Statistical Genomics, Division of Intramural Research Program,
National Institute of Mental Health, Bethesda, MD, USA
e-mail: kay1yao@mail.nih.gov

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_11

The shift to linkage was important because this analysis uses genetic markers. These genetic markers can be identified as potential risk factors for the disorder being studied. This analysis, like segregation, requires the use of family data. Linkage analyses are able to identify regions of these markers, but much of the genome is missing in the analysis. Linkage is named so because genes that are physically close to one another on a chromosome remain linked during meiosis (Pulst 1999). The results of this analysis are region specific and do not have the specificity of identifying at the single nucleotide polymorphism (SNP) level. The available genetic markers are fewer in number than at the genome-wide association analysis (GWAS). Thus, linkage analyses shifted toward using GWAS.

Using the GWAS method is preferable to the previous three mentioned because it does not require the use of family data. Data can be in the form of cases and controls, meaning the cases are individuals affected by the disorder, and controls are individuals without the disorder. With the switch to GWAS also came a complete genome. As mentioned, linkage analyses did not have the ability to utilize genetic markers from the entire genome, while GWAS is able to get a clearer picture. GWAS's seek to determine if specific SNPs are associated with a trait/phenotype of a disorder (Bush and Moore 2012). The GWAS method of analysis is still widely used, especially in the realm of psychiatric disorders (Ripke et al. (2017); Power et al. 2017; Ritter et al. 2017). Within the use of GWAS, meta-analyses are also being used to pool GWAS data from multiple analyses and analyze it collectively. The progression of these molecular frameworks can be seen in Fig. 11.1. This

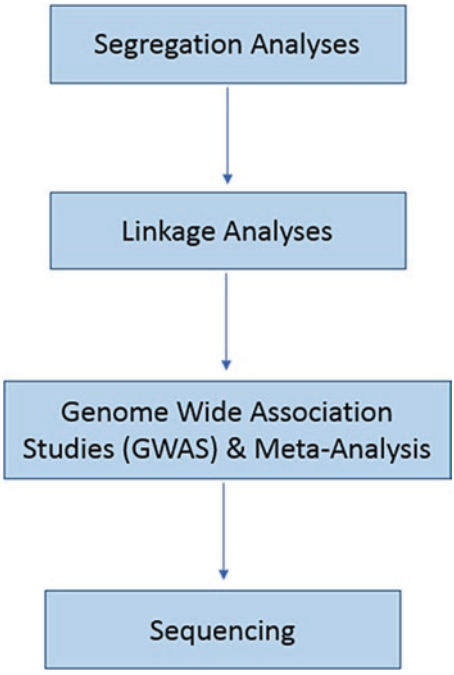


Fig. 11.1 The progression of molecular frameworks

chapter will discuss each of the above mentioned statistical methods in the context of OCD, as well as provide future insights to the direction of OCD analyses.

11.2 Segregation Analyses

Segregation analyses seek to determine the pattern of inheritance of the disorder by analyzing family data. Complex segregation analyses can be conducted using the S.A.G.E. software (Case Western 1997). A regressive logistic model is used and various variables of interest are included (i.e., age of onset). The developed model then tests for the presence of a major susceptibility locus. Several Mendelian models are used to try and fit the data to determine the mode of inheritance of the disorder. These models usually include a Mendelian codominant model, Mendelian dominant model, Mendelian recessive model, and a Mendelian additive model (Hanna et al. 2005a). The models of unrestricted and no mode of transmission are also commonly included in the analysis.

Several segregation analyses were conducted for OCD. Hanna et al. found that the Mendelian dominant model best fit the data (2005A). It was also documented that among families with OCD, there was evidence of a major susceptibility locus when age of onset was included in the model (Hanna et al. 2005a). Previous to this study, age of onset had not been used in OCD models. This analysis used pediatric probands, and with the combined use of age of onset, it is possible they contributed to the detection of a major effect. This is because of the high probability of OCD occurring in relatives of the probands that presented with an early age of onset of OCD (Pauls et al. 1995; Nestadt et al. 2000a; Hanna et al. 2005b).

Another segregation analysis conducted by Cavallini et al. found too that the best fit model was the dominant mode of transmission for Cavallini et al. (1999). Even though the dominant model was supported by the data, there was potential for the presence of genetic heterogeneity (Cavallini et al. 1999). Two different phenotypes were used, which included: (1) OCD; (2) OCD plus Tourette's syndrome (TS)/chronic motor tics (CMT). When only OCD families were used, the dominant model was best, but when the phenotype was extended to include OCD, TS, and CMT, the unrestricted model of transmission was found to be the best fit. It was found that there was recurrence of TS in OCD families, as well as OCD in TS families, but there was not enough evidence to determine the existence of a bidirectional pathway between them. This suggested that OCD is clinically heterogeneous and further research should be done to examine the phenotypes of the disorder.

Lastly, a segregation analysis conducted by Nestadt et al. also found that the Mendelian dominant model was the best fit of the OCD data (2000B). There was evidence of heterogeneity of the families separated by male and female probands. Because of this, separate segregation analyses were conducted separated by sex. The families with a male proband found that a Mendelian major-locus model was supported, but the specific type of Mendelian model was unclear because each of them fit as well as the general model (Nestadt et al. 2000b). Using the families with

the female proband, the results were similar to that of the total sample. Neither the dominant nor codominant model could be rejected, but using the AIC, the dominant model was chosen (Nestadt et al. 2000b).

Segregation analyses are beneficial in that they can help determine which mode of transmission is most likely for a disease, but they cannot identify genetic markers that may contribute to this heritability. Segregation can only say that there is evidence of a major susceptibility locus or gene for that mode of transmission. The transition to linkage analyses helped to specify heritability in that regions of genetic markers could be identified as contributing to the inheritance of that disorder.

11.3 Linkage Analyses

Linkage analyses are able to identify regions of markers that may contribute to the heritability of a certain disorder. Most commonly, Morgan (http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml) and Merlin (http://www.sph.umich.edu/csg/abecasis/Merlin/index.html) software are used to complete these analyses. Morgan is able to analyze very large pedigrees, but the number of markers is limited and they must all be in linkage equilibrium (Mathews et al. 2012). Merlin controls for the effects of linkage disequilibrium between the markers, but all individuals cannot be used because of both the size and complexity of large families (Mathews et al. 2012). The regions are measured based on a heterogeneity logarithm of odds score (HLOD), which estimates whether two loci are located close enough together on a chromosome to be inherited together. An HLOD score of ≥ 1.5 is a threshold commonly used to identify an area of interest.

Mathews et al. conducted a linkage analysis on families affected by OCD and found 11 chromosomal regions with HLOD scores ≥ 1.5 and 5 with a HLOD score >2 (2012). Chromosome 1p36 contained a region with the highest HLOD scores of 2.96 under the dominant model using Merlin and 2.66 under the dominant model using Morgan (Mathews et al. 2012). The linkage region from 1p35.33 to 1p36.32 met genome-wide criteria for suggestive linkage, which was the strongest finding. The 1p36 region has been implicated in a deletion syndrome (1p36 syndrome) that is characterized by intellectual disability, as well as multiple system abnormalities (Battaglia et al. 2008).

Another linkage analysis conducted on Costa Rican families found the strongest LOD score on chromosome 15q14 with a score of 3.13 (Ross et al. 2011). Eleven chromosomal regions had LOD scores ≥ 1.5 and four with a score >2 (Ross et al. 2011). Chromosome 15q14 was previously implicated in studies of compulsive behavior in mice (Kas et al. 2010). Shugart et al. also implicated 15q14 as a region of interest for Shugart et al. (2006).

The Shugart et al. linkage analysis used Kong and Cox logarithm of odds scores (KAC LOD) (2006). The strongest signal was found on chromosome 3 at D3S1262 with a KAC LOD of 2.67 (Shugart et al. 2006). Suggestive linkage signals were also

found on chromosomes 3q27-28, 6q, 7p, 1q, and 15q. As mentioned, the 15q chromosome has been linked to compulsive behavior in mice (Kas et al. 2010).

Linkage analyses are able to identify regions of markers, but genome-wide association studies (GWAS) can isolate specific SNPs that can be linked back to genes that may contribute to the susceptibility of a disorder. The use of GWAS resulted because of increased specificity of the results.

11.4 Genome-Wide Association Studies and Meta-analysis

Genome-wide association studies (GWAS) seek to establish an association between a specific disorder and genetic variants to determine which variants affect susceptibility of that disease. The gold standard for the genome-wide significance level is 5.0×10^{-8}. PLINK (http://zzz.bwh.harvard.edu//plink/) is used for quality control to remove poorly genotyped SNPs as well as individuals as a whole. PBAT (https://www.hsph.harvard.edu/pbat/download2/) is then used to conduct the association analysis when dealing with complex family structures. A meta-analysis can also be included with a GWAS if several results combined and then analyzed together. METAL (http://csg.sph.umich.edu/abecasis/metal/download/) is the software used to conduct the meta-analysis.

A GWAS and meta-analysis were completed on the OCD Collaborative Genetics Association Study (OCGAS) data (Mattheisen et al. 2015). After quality control via PLINK, PBAT was used to compute p-values for autosomal markers and the within and between family information (Mattheisen et al. 2015). These computed p-values were combined and analyzed in METAL for the meta-analysis. None of the SNPs reached genome-wide significance. The most significant SNP was rs4401971, which is located near the PTPRD gene, with a p-value of 4.13×10^{-7} (Mattheisen et al. 2015). Pre-synaptic PTPRD promotes the differentiation of glutamatergic synapses (Dunah et al. 2005; Woo et al. 2009; Kwon et al. 2010; Takahashi et al. 2011). It also interacts with Slit and NTRK-like family member 3 (SLITRK3), which is a postsynaptic adhesion molecule. Molecules in the same family (SLITRK5 and SLITRK1) have been associated with TS as well as OCD (Abelson et al. 2005; Shmelkov et al. 2010). This study had a large overlap of signals with the IOCDF-GC study (Stewart et al. 2013). Twelve of the 15 strongest signals from the Stewart et al. study overlapped with the results of the Mattheisen et al. study (Stewart et al. 2013; Mattheisen et al. 2015).

The Stewart et al. study used trio data and the most significant SNP was rs6131295 with a p-value of 3.84×10^{-8} (2013). The top two SNPs from the case control meta-analysis (rs6131295 and rs10165908) are found within the DLGAP1 genes, which influences glutamate signaling. Several of the other top SNPs from the combined trio-case-control meta-analysis are found in or near genes that have been implicated with other psychiatric disorders, which include ADCY8 (Potkin et al. 2009; de Mooij-van Malsen et al. 2009; Kantojarvi et al. 2010), ARHGAP18 (International

HapMap C 2005), and JMJD2C (Rivière et al. 2009) in bipolar disorder, schizophrenia, and autism spectrum disorders, respectively (Stewart et al. 2013).

GWAS and meta-analyses are still widely used in the field. All of the previous analyses have focused on common variants, and the hope for the future is to continue sequencing in order to have access to rare variants for further analysis.

11.5 Future Looks

Current OCD genetic analyses utilize common variants and the next step will be to switch to analyzing rare variants. Additionally, the use of larger sample sizes is needed to identify many of the rare variants. We expect that the contribution of rare variants, as well as larger samples sizes, will help to further identify risk factors for OCD, as well as other disorders. A meta-analysis combining two OCD consortiums (International Obsessive Compulsive Disorder Foundation Genetics Collaboration (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS)) found that even though a large sample size was used through the combination of the two datasets, the study was still underpowered. Thus, both national and international collaborations are warranted.

Phenotypically, more funding should be placed in research efforts to collect data on OCD cases that contain more precise and accurate phenotypes, given that the disorder is clinically complex. OCD is both genetically and clinically heterogeneous, which makes defining a group of cases all as having strict OCD difficult. The heterogeneity of OCD may play a role in both the underpowered results and a lack of signals showing up in GWAS and meta-analyses.

With further research efforts focusing on both sequencing of rare variants and collecting OCD cases with precisely defined phenotypes, the hope is to find more concrete genetic markers that contribute to the susceptibility of OCD.

References

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, et al. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. 2005;310(5746):317–20. [PubMed: 16224024]

Battaglia A, Hoyme HE, Dallapiccola B, Zackai E, Hudgins L, McDonald-McGinn D, et al. Further delineation of deletion 1p36 syndrome in 60 patients: a recognizable phenotype and common cause of developmental delay and mental retardation. Pediatrics. 2008;121:404–10.

Bokor G, Anderson P. Obsessive-compulsive disorder. J Pharm Pract. 2014;27(2):116–30. https://doi.org/10.1177/0897190014521996.

Bush WS, Moore JH. Chapter 11: genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822. https://doi.org/10.1371/journal.pcbi.1002822.

Cavallini MC, Psaquale L, Bellodi L, Smeraldi E. Complex segegation analysis for obsessive compulsive disorder and related disorders. Am J Med Genet Neuropsychiatr Genet. 1999;88:38–43.

Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al. Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS Genet. 2013;9(10):e1003864. https://doi.org/10.1371/journal.pgen.1003864.

de Mooij-van Malsen AJ, van Lith HA, Oppelaar H, Hendriks J, de Wit M, Kostrzewa E, et al. Interspecies trait genetics reveals association of Adcy8 with mouse avoidance behavior and a human mood disorder. Biol Psychiatry. 2009;66(12):1123–30. [PubMed: 19691954]

Dunah AW, Hueske E, Wyszynski M, Hoogenraad CC, Jaworski J, Pak DT, et al. LAR receptor protein tyrosine phosphatases in the development and maintenance of excitatory synapses. Nat Neurosci. 2005;8(4):458–67. [PubMed: 15750591]

Elston RC. Segregation analysis. In: Harris H, Hirschhorn K, editors. Advances in human genetics 11. Boston: Springer; 1981.

Hanna GL, Fingerlin TE, Himle JA, Boehnke M. Complex segregation analysis of obsessive-compulsive disorder in families with pediatric probands. Hum Hered. 2005a;60(1):1–9. https://doi.org/10.1159/000087135.

Hanna GL, Himle JA, Curtis GC, Gillespie BW. A family study of obsessive-compulsive disorder with pediatric probands. Am J Med Genet Neuropsychiatr Genet. 2005b;134B:13–9.

Kantojärvi K, Onkamo P, Vanhala R, Alen R, Hedman M, Sajantila A, et al. Analysis of 9p24 and 11p12-13 regions in autism spectrum disorders: rs1340513 in the JMJD2C gene is associated with ASDs in Finnish sample. Psychiatr Genet. 2010;20(3):102–8. [PubMed: 20410850]

Kas MJ, Gelegen C, van Nieuwerburgh F, Westenberg HG, Deforce D, Denys D. Compulsivity in mouse strains homologous with chromosomes 7p and 15q linked to obsessive-compulsive disorder. Am J Med Genet B Neuropsychiatr Genet. 2010;153B(1):252–9. [PubMed: 19514050]

Katerberg H, Delucchi KL, Stewart SE, Lochner C, Denys DA, Stack DE, et al. Symptom dimensions in OCD: item-level factor analysis and heritability estimates. Behav Genet. 2010;40(4):505–17. https://doi.org/10.1007/s10519-010-9339-z.

Kwon SK, Woo J, Kim SY, Kim H, Kim E. Trans-synaptic adhesions between netrin-G ligand-3 (NGL-3) and receptor tyrosine phosphatases LAR, protein-tyrosine phosphatase delta (PTPdelta), and PTPsigma via specific domains regulate excitatory synapse formation. J Biol Chem. 2010;285(18):13966–78. [PubMed: 20139422]

Mathews CA, Badner JA, Andresen JM, Sheppard B, Himle JA, Grant JE, et al. Genome-wide linkage analysis of obsessive-compulsive disorder implicates chromosome 1p36. Biol Psychiatry. 2012;72(8):629–36. https://doi.org/10.1016/j.biopsych.2012.03.037.

Mattheisen M, Samuels JF, Wang Y, et al. Genome-wide association study in obsessive compulsive disorder: results from the OCGAS. Mol Psychiatry. 2015;20:337–44. [PubMed: 24821223]

Nestadt G, Samuels J, Riddle M, Bienvenu OJ III, Liang K-Y, LaBuda M, Walkup J, Grados M, Hoehn-Saric R. A family study of obsessive-compulsive disorder. Arch Gen Psychiatry. 2000a;57:358–63.

Nestadt G, Lan T, Samuels J, Riddle M, Bienvenu O, Liang K, et al. Complex segregation analysis provides compelling evidence for a major gene underlying obsessive-compulsive disorder and for heterogeneity by sex. Am J Hum Genet. 2000b;67(6):1611–6. https://doi.org/10.1086/316898.

Pauls DL, Alsobrook JP II, Goodman W, Rasmussen S, Leckman JF. A family study of obsessive-compulsive disorder. Am J Psychiatry. 1995;152:76–84.

Potkin SG, Turner JA, Fallon JA, Lakatos A, Keator DB, Guffanti G, et al. Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia. Mol Psychiatry. 2009;14(4):416–28. [PubMed: 19065146]

Power RA, Tansey KE, Buttenschon HN, Cohen-Woods S, Bigdeli T, Hall LS, et al. Genome-wide association for major depression through age at onset stratification: major depressive disorder working group of the psychiatric genomics consortium. Biol Psychiatry. 2017;81:325–35.

Pulst SM. Genetic linkage analysis. Arch Neurol. 1999;56(6):667–72. https://doi.org/10.1001/archneur.56.6.667.

Ripke S, Group SW, O'Donovan M. Current status of schizophrenia GWAS. Eur Neuropsychopharmacol. 2017;27:S415. https://doi.org/10.1016/j.euroneuro.2016.09.460.
Ritter ML, Guo W, Samuels JF, Wang Y, Nestadt PS, Krasnow J, et al. Genome wide association study (GWAS) between attention deficit hyperactivity disorder (ADHD) and obsessive compulsive disorder (OCD). Front Mol Neurosci. 2017;10(83). https://doi.org/10.3389/fnmol.2017.00083.
Rivière JB, Xiong L, Levchenko A, St-Onge J, Gaspar C, Dion Y, et al. Association of intronic variants of the BTBD9 gene with Tourette syndrome. Arch Neurol. 2009;66(10):1267–72. [PubMed: 19822783]
Ross J, Badner J, Garrido H, Sheppard B, Chavira DA, Grados M, et al. Genomewide linkage analysis in Costa Rican families implicates chromosome 15q14 as a candidate region for OCD. Hum Genet. 2011;130(6):795–805. https://doi.org/10.1007/s00439-011-1033-6.
S.A.G.E. (Statistical Analysis for Genetic Epidemiology), version 3.1. Computer program package, available from the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland. 1997.
Shmelkov SV, Hormigo A, Jing D, Proenca CC, Bath KG, Milde T, et al. Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive-like behaviors in mice. Nat Med. 2010;16(5):598–602. 591p following 602
Shugart YY, Samuels J, Willour VL, Grados MA, Greenberg BD, Knowles JA, et al. Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q and 6q. Mol Psychiatry. 2006;11(8):763–70. https://doi.org/10.1038/sj.mp.4001847.
Stewart SE, Platko J, Fagerness J, Birns J, Jenike E, Smoller JW, et al. A genetic family-based association study of OLIG2 in obsessive-compulsive disorder. Arch Gen Psychiatry. 2013;64(2):209–14. [PubMed: 17283288]
Takahashi H, Arstikaitis P, Prasad T, Bartlett TE, Wang YT, Murphy TH, et al. Postsynaptic TrkC and presynaptic PTPsigma function as a bidirectional excitatory synaptic organizing complex. Neuron. 2011;69(2):287–303. [PubMed: 21262467]
The International HapMap C. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320. [PubMed: 16255080]
Woo J, Kwon SK, Choi S, Kim S, Lee JR, Dunah AW, et al. Trans-synaptic adhesion between NGL-3 and LAR regulates the formation of excitatory synapses. Nat Neurosci. 2009;12(4):428–37. [PubMed: 19252495]

Web Resources

http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml
http://www.sph.umich.edu/csg/abecasis/Merlin/index.html
http://zzz.bwh.harvard.edu//plink/
https://www.hsph.harvard.edu/pbat/download2/
http://csg.sph.umich.edu/abecasis/metal/download/