Title

Human genetic-epidemiologic association analysis via allelic composition and DNA sequence similarity methods : applications to blood-based gene expression biomarkers of disease

Author
Wessel, Jennifer

Publication Date
2006

UNIVERSITY OF CALIFORNIA, SAN DIEGO

SAN DIEGO STATE UNIVERSITY

Human Genetic-Epidemiologic Association Analysis via Allelic Composition and DNA Sequence Similarity Methods: Applications to Blood-Based Gene Expression Biomarkers of Disease.

A dissertation submitted in partial satisfaction of the requirements

for the degree of Doctor of Philosophy

in Public Health (Epidemiology)

by

Jennifer Wessel

Committee in charge:

University of California, San Diego

Professor Nicholas Schork, Chair
Professor Sonia Jain
Professor Daniel O'Connor
Professor Deborah Wingard

San Diego State University

Professor Ming Ji
Professor Jeanette McCarthy

2006

The dissertation of Jennifer Wessel is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

____
Chair

University of California, San Diego

San Diego State University

2006

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2 **Page**

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF TABLES

CHAPTER 1 **Page**

CHAPTER 2

CHAPTER 3

CHAPTER 4

ACKNOWLEDGEMENTS

First and foremost I have to thank Nik Schork for giving me the opportunity and believing in me. His positive attitude gave me the strength and wherewithal to keep trying. In addition, his excellent ideas stimulated the basis of my research. I also have to thank Jeanette McCarthy for teaching me a strong foundation in genetic epidemiology. She was very hands on with me, which as a new student of genetic epidemiology I needed to get started. I want to thank Debbie Wingard for her thoughtful critiques and feedback, as well as her guidance in the principles and understanding of the practice of epidemiology. Daniel O'Connor spent a lot of time with me and I appreciate his teaching and thorough understanding of the physiology of hypertension and the kidney. Sonia Jain has always been approachable, thoughtful and conscientious – she always gave me well thought out feedback when I needed it. Ming Ji, who I have worked with often and throughout my education, has taught me a great deal about statistics.

While my professors were instrumental in my education, the support of my family and friends has brought me this far. I have to thank Vernon Franck for sticking it through with me; he supported me through the hard times. Jochen Schmitz understood the difficulties of going through graduate school and gave me the support and motivation in those last few years to finish it. My sister, Kristin Wessel, always sympathized with me through all of my struggles. The birth of my nephew, Simon Wessel, brought hope and reason to it all. And of course my mother, Jacqueline Wessel, who raised us on her own, has always

been supportive and positive about it all and most importantly taught me to keep going even when it wasn't easy.

The text of Chapter 2 has been published in the American Journal of Human Genetics (79(5):792-806, November, 2006) for publication as:

> Jennifer Wessel, and Nicholas J Schork. Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis

The text of Chapters 3 and 4 have been submitted for publication as:

> Jennifer Wessel, Ondrej Libiger, and Nicholas J. Schork. Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis.
>
> Jennifer Wessel, Matt Zapala and Nicholas J. Schork. Accommodating Pathway Information in Expression Quantitative Trait Locus ("eQTL") Analysis.

VITA

EDUCATION

1996	BS, Molecular & Cellular Biology, University of Arizona
1999	Master of Public Health, Epidemiology, University of Arizona
2006	Doctor of Philosophy, University of California, San Diego

PROFESSIONAL EXPERIENCE

2003 – present	Graduate researcher, University of California, San Diego, Department of Psychiatry, Nicholas J Schork.
2001 – 2003	Graduate researcher, San Diego State University, Graduate School of Public Health, Jeanette McCarthy.
2000 – 2001	Graduate researcher, Naval Health Research Center, Jerry Larson.

PUBLICATIONS

2006

Jennifer Wessel*, Guillermo Moratorio*, Fangwen Rao, William Greene, Brinda K. Rana, Brian P. Kennedy, Elizabeth O. Lillie, Douglas W. Smith, Michael G. Ziegler, Nicholas J. Schork, Geert W. Schmid-Schönbein, and Daniel T. O'Connor. C reactive protein, an "intermediate phenotype" for inflammation: human twin studies reveal its heritability, association with blood pressure and the metabolic syndrome, and the influence of polymorphism at adrenergic pathway loci. Journal of Hypertension, Accepted October 2, 2006.

J Wessel, NJ Schork. Genomic Distance-Based Regression Analysis for Testing Multilocus Associations with Quantitative and Qualitative Phenotypes. American Journal of Human Genetics, 2006 Nov;79(5):792-806.

Yan Gong, Amber L. Beitelshees, Jennifer Wessel, Taimour Y. Langaee, Nicholas J. Schork, Julie A. Johnson. SNP Discovery and Haplotype Analysis of BK Channel Beta1 Subunit. Accepted Aug 30, 2006. Pharmacogenetics and Genomics.

Carol A Mathews, Caroline M Nievergelt, Amin Azzam, Helena Garrido, Denise A Chavira, Jennifer Wessel, Monica Bagnarello, Victor I Reus, Nicholas J Schork. Heritability and Clinical Features of Multigenerational Families With Obsessive-

Compulsive Disorder and Hoarding. Accepted May 4, 2006. Neuropsychiatric Genetics.

Jennifer Wessel*, Tammy M. Seasholtz*, Fangwen Rao, Brinda K. Rana, Brian P. Kennedy, Michael G. Ziegler, Douglas W. Smith, Nicholas J. Schork, Joan H. Brown, Daniel T. O'Connor. Rho kinase polymorphism influences blood pressure and systemic vascular resistance in human twins: role of heredity. Hypertension, 2006 May;47(5):937-47.

2005

Salem, Rany M.; Wessel, Jennifer; Schork, Nicholas J.;. A Comprehensive Literature Review of Haplotyping Software and Methods for Use with Unrelated Individuals. *Human Genomics* 2, 39-66 (2005). Invited Review. http://polymorphism.ucsd.edu/HapSoftwareReview/

2004

Lian Zhang, Fangwen Rao, Jennifer Wessel, Brian P. Kennedy, Brinda K. Rana, Laurent Taupenot, Elizabeth O. Lillie, Nicholas J. Schork, Michael G Ziegler, and Daniel T O'Connor. (2004) Functional Allelic Heterogeneity and Pleiotropy of a Repeat Polymorphism in Tyrosine Hydroxylase: Prediction of Catecholamines and Response to Stress in Twins. Physiological Genomics. 2004 Nov 17;19(3):277-291.

J Wessel, JJ McCarthy, EJ Topol. (2004) Replication of the Association Between the Thrombospondin-4 A387P Polymorphism and Myocardial Infarction: Importance of Waist-to-hip Ratio As a Confounding Variable.   American Heart Journal . 2004 May;147(5):905-9.

Papers Submitted.

Jennifer Wessel, Ondrej Libiger, and Nicholas J. Schork. Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis. Submitted to Genetic Epidemiology

Jennifer Wessel, Matt Zapala and Nicholas J. Schork. Accommodating Pathway Information in Expression Quantitative Trait Locus ("eQTL") Analysis. Submitted to Genomics.

Jennifer Wessel, Andrew J. Schork, and Nicholas J. Schork. Powerful designs for gene-phenotype associations that exploit related individuals. Submitted to Genetic Epidemiology.

C Ye, MA Zapala, HM, Kang, J Wessel, E Eskin, N Schork. High-Density QTL Mapping to Identify Phenotypes and Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways. Submitted to Molecular Systems Biology.

G Wen, J Wessel, B Hamilton, NJ Schork, DT O'Connor. Discovery of polymorphisms in secretogranin-2 and their effect on blood pressure phenotypes. Submitted Journal of Clinical Investigation.

Su C-Y, Corby PM, Elliott MA, Studen-Pavlovich DA, Ranalli DN, Rosa B, Wessel J; Schork NJ, Hart TC, Bretz WA. Inheritance of Occlusal Topography: A Twin Study. Submitted Pediatric Dentistry.

Fangwen Rao, Jennifer Wessel, Gen Wen, Lian Zhang, Brinda K. Rana, Brian P. Kennedy, Elizabeth O. Lillie, R Salem, Y Chen, B Hamilton, D Smith, N Holstein-Rathlou, M Ziegler, NJ Schork, and DT O'Connor. Renal microalbumin excretion, an "intermediate phenotype" for nephropathy: Human twin studies identify the influences of heredity, environment, and the interactive role of polymorphism at adrenergic loci. Submitted Journal of the American Society of Nephrology.

Fangwen Rao, Lian Zhang, L Taupenot, Gen Wen, Jennifer Wessel, Brian P. Kennedy, K Zhnag, Brinda K. Rana, DW Smith, EO. Lillie, PE Cadman, R Salem, NJ Schork, M Ziegler,and DT O'Connor. Tyrosine hydroxylase (TH), the rate-limiting enzyme in catecholamine biosynthesis: Discovery of common human genetic variants governing transcription, human autonomic activity and blood pressure in vivo. Submitted Circulation.

J Wessel, D Kritz-Silverstein, D Morton, D Wingard, E Barrett-Connor. ApoE4 and Heart Disease Risk Factors at Midlife Influence Cognitive Function in Older Women and Men: The Rancho Bernardo Study.

Papers in Process.

Wessel J, Zapala M, Schork NJ. Applications to Blood-based Gene Expression Bio-markers of Disease.

Wessel, Jennifer; Salem, Rany M.; Schork, Nicholas J.; A Comprehensive Literature Review of Haplotyping Software and Methods for Use with Pedigrees and Related Individuals. *Human Genomics* (2006). Invited Review. http://polymorphism.ucsd.edu/HapSoftwareReview/

ABSTRACT OF THE DISSERTATION

Human Genetic-Epidemiologic Association Analysis via Allelic Composition and DNA Sequence Similarity Methods: Applications to Blood-Based Gene Expression Biomarkers of Disease.

by

Jennifer Wessel

Doctor of Philosophy in Public Health (Epidemiology)

University of California, San Diego, 2006

San Diego State University, 2006

Professor Nicholas J Schork, Chair

The Human Genome Project, and related DNA sequence variation projects, has provided researchers with both the motivation and raw material for considering large-scale genetic association studies seeking to identify genetic variations that contribute to disease susceptibility. Association studies are plagued by many problems, including inappropriate data analysis methodologies and the potential for false positive results due to the testing of hundreds-of-thousands, of polymorphic loci for association with a disease. Most association analysis methodologies ignore biological realities mediating gene-phenotype relationships, such as the possibility that genes and genetic variations work in concert or in combination to influence a disease and/or phenotypic expression. I describe a statistical analysis methodology for association studies which considers the genetic variation within a gene (chapter 2), across the entire genome (chapter 3), or a series of genes in pathways (chapter 4), as “wholes” rather than as individual isolated entities that are to be assessed independently of each other. I showcase the methodology by applying it to publicly available genotype and gene expression data from the HapMap Project on 57 CEPH individuals. I provide biological motivation for this type of analysis approach and consider measures that assess the “genomic similarity” of individuals with respect to the variations they possess across a number of loci. I, describe a weighted distance-based regression method that exploits this similarity measure in association analyses. In chapter 2, I develop and apply the method to an analysis of the CHI3L2 gene and document the utility and flexibility of the method. In chapter 3, I apply the method developed in chapter 2 to a whole genome analysis of 811,886 phased genetic variations typed on the CEPH subjects. In chapter 4, I extend the method to the analysis of biochemical pathways involved in diseases, functions, and drug targets that are affected by multiple SNPs. I ultimately argue that my work has the potential to not only open up a new area of research in genetic epidemiology and statistical genetic methodology, but also to shed light on the genetic basis of complex, multifactorial diseases and phenotypes.

CHAPTER 1

Introduction and Background

INTRODUCTION AND BACKGROUND

The Emergence of Genetic Epidemiology

The research disciplines of genetic and molecular epidemiology have grown considerably in the last few decades as a result of the Human Genome Project [1, 2], the International HapMap Project [3], and related initiatives in human genetics, as well as the very recent development of technologies such as multiplex, microarray assays and other high-throughput molecular phenotyping technologies [4-8]. A few recent reviews have provided overviews of the motivations and goals of the field of genetic epidemiology arguing that genetic epidemiology deals with the etiology, distribution, and control of disease in groups of relatives and/or with respect to the inherited causes of disease in population in the population at large [9-11]. In fact, many very recent efforts have been made to incorporate the increasingly sophisticated understanding of the human genome into public health and epidemiology research and practice. For example, the National Office of Public Health Genomics at the Centers for Disease Control disseminates information on how genomic discoveries can be used to improve health and prevent disease (http://www.cdc.gov/genomics/default.htm).

With current technologies it is now possible to collect massive amounts of genetic information on individuals sampled for very large-scale epidemiologic studies. Making sense of this information in order to, e.g., identify inherited variations that contribute to disease susceptibility, or identify molecular biomarkers of disease progression, is complicated given the number of variables that might need to be considered in relevant statistical analyses of the data. In addition, sorting out the biological meaningfulness of the results of the statistical analyses is itself enormously complicated.

My thesis research has focused on the development of data analysis techniques appropriate for relating genetic variation to phenotypic variation that are applicable to large-scale genetic epidemiologic studies [12, 13]. I have applied these analysis techniques to blood and immortalized lymphocyte-based gene expression data obtained on individuals that have been genotyped on a large panel of genetic markers. Gene expression “fingerprints” obtained from accessible tissues such as blood are growing in popularity as providing diagnostic and prognostic biomarkers for clinical and epidemiologic studies of disease [14-16]. Thus, my research encompasses issues in genetic and molecular epidemiology, epidemiologic biomarker analysis, data analysis, and integrated approaches to understanding the biological basis of human disease from population-level, epidemiologic analysis perspectives.

Modern Genomics and Association Studies

The search for genetic variations that contribute to complex, multifactorial traits and diseases, such as hypertension or cancer, has been problematic[12]. The reasons for this are somewhat obvious, in that the influence and identification of each particular gene or environmental factor that impacts the expression of such traits and diseases are often obscured or confounded by the effects of the other factors. As a result, studies that test the association of a particular mutation or genetic variant to a trait or disease with the hope of identifying a factor contributing to the expression of that trait or disease do not often yield compelling, replicable results [12]. To overcome this problem, at least in part, The International HapMap Project (IHP) [17] – considered the “second generation” Human Genome Project – was initiated, seeking to identify the smallest set of genome-wide genetic markers that is likely to yield the greatest probability of identifying a gene in studies designed to test each and everyone of those markers for association with a trait or disease. The initiation of the IHP has led to developments in high throughput sequencing and genotyping (e.g. “SNP chips” where 500,000 Single Nucleotide Polymorphisms (SNPs) can be interrogated on s ingle individual in a single assay are available and soon will be capable of interrogating 1,000,000 SNPs), to facilitate the massive amounts of genotyping that might be necessary to implement genome-wide association studies of the type envisioned by the IHP. In fact, recent multi-million dollar research initiatives announced by the National Institutes of Health, such as the Genetic Association Information Network (GAIN) initiative (http://www.fnih.org/GAIN/GAIN_home.shtml) and the Gene x Environment Interaction (GEI) initiative (http://www.genome.gov/17516707) attest to the emphasis being placed on large-scale association studies in biomedical and epidemiologic research.

A number of very recent scientific articles and study reviews have appeared in the last few years that discuss the design and analysis of data generated by genome-wide studies, including those that make use of IHP data[18,19]. While the IHP seeks to minimize the number of marker loci one needs to effectively interrogate the whole genome for association with a trait or disease, its basic design, which is rooted in the belief that one can effectively capture important “functionally significant” variants through evolutionary and/or historical linkage disequilibrium (LD) relationships with neighboring loci captured by haplotype analyses ignores five fundamental facts about the human genome and human physiology: 1. Humans are diploid; 2. Genetic variation within genes is not likely to act in isolation but rather manifests more “holistic” phenotypic effects such that studying individual loci might not capture the effect of the variations; 3. The evolutionary history of genes may not match the biological functions of those genes, nor the variations within them, such that determining what an individual inherited from each parent (i.e., haplotyping) might not be as important an issue as determining, in great detail, simply what set of variants an individual possesses in his or her genome; 4. At the sequence level, each individual may have a unique set of variations if a large enough region is studied; 5. Studies investigating the *in vitro* and *in silico* functional significance of genes and genetic variations are being pursued on a large-scale. Each of these items is discussed in more detail in Chapter 2.

The reason these five factors are important to consider in association studies is rooted in the way individual genomes are constructed during meiosis and how their construction from parental genomes dictates the unique phenotypic expression pattern exhibited by every individual. Roughly 10 million sites within the human genome have been shown to be “polymorphic” (i.e., vary from person to person). Most of these variations are single nucleotide polymorphisms or “SNPs,” of which at least 5 million have been described in public databases [20, 21, 22]. SNPs that are closely linked (i.e., physically close to each other) are not inherited independently from each other due to the fact that recombination during meiosis does not occur so frequently as to disrupt the transmission of large segments of parental chromosomes intact to offspring. Thus, the closer two loci are on the genome, the less likely a recombination event is to occur between them, thus allowing variations at those loci to be transmitted together on a single chromosomal segment. This phenomena creates associations between variations at neighboring genomic positions in the population at large and is the phenomena that the IHP is trying to exploit in developing a “haplotype” or linkage disequilibrium (LD) map of the genome[23, 24].

A major challenge in the study of genetic variation – which is seen as one of the motivating factors for the IHP, is thus the determination of the subset of the 10 million polymorphic sites that harbor functional variations and are thus associated with particular phenotypes (as opposed to variations that are merely in linkage disequilibrium with the functional variations), and how such functional variation contributes to phenotypic expression.

In this post-human genome and post-IHP era, many thousands of SNPs are being tested for their association with complex, common diseases. The results from these studies have often been inconclusive and controversial. As a result, many researchers have begun to consider the reasons why such studies are so problematic. Several reasons have been suggested. These include 1. the enormous statistically-challenging multiple comparisons problem that arises when one considers thousands, if not hundreds of thousands, of markers for association with a trait or disease; 2. the weak effects of each gene and lack of statistical power in samples studied to date; 3. unexplored gene-environment interactions; 4. the determination of a plausible biological context within which an associated gene can be proven to influence phenotypic expression; 5. population stratification and other factors that may confound statistical tests; 6. population differences in linkage disequilibrium; 7. disease heterogeneity; and 8. general study design issues (e.g. phenotypic heterogeneity, sampling of controls, etc.) [12, 25-35]. A very recent review examined all known genetic association studies and suggested that of the over 600 positive associations reported in the literature between a genetic variation and a trait or disease, 166 had been studied three or more times, and only 6 of those have been consistently replicated [12].

Limitations in Contemporary Analysis Methods for Association Studies

Of the eight or so factors that could create problems for large-scale genome-wide association studies described in the previous section, those that concern data analysis are particularly thorny. The most basic approach to the analysis of data generated as part of a whole-genome association study of the type envisioned by the IHP is to test each individual locus for association with the trait or disease in question independently of the other loci. This assumes that the effects of each locus, both within and across genes, on phenotypic expression are independent. Although there is some research that considers the analysis of interactions between or *across* different genes for association studies [36], there is little research that considers the simultaneous effect of multiple variations *within* a gene. Thus, an alternative or complementary analysis approach to genetic associations would involve consideration of the actual *composition* of genes (i.e., consideration of the effects of particular combinations of variations in a gene that an individual possesses), and the impact that these multiple variations have on phenotypic expression. The primary aim of this dissertation is to develop and apply methods of analysis that would address questions of *why* one should consider such approaches to genetic association analysis as well as *how* one can construct relevant analytical methods to implement these approaches.

Molecular Phenotyping and Biomarker Analysis

Increasing emphasis in epidemiologic studies of disease susceptibility is being placed on subclinical phenotypes and/or biomarkers of disease (see Tables 1-3 for references). Emerging technologies such as gene expression, proteomic, metabolomic, and imaging technologies have revolutionized the way epidemiologists can consider assaying a disease and disease process. Accessible tissue-based microarray-oriented gene expression analysis, in particular, has been pursued by a number of investigators with respect to a wide-variety of diseases (see Table 3 for references).

Patterns in gene expression data have been used by researchers to understand how an exposure or a disease process may influence, e.g., gene expression patterns, protein levels, interactions among gene, or the subclinical physiologic state of individuals in general. I have participated in a review of the literature for studies using blood to investigate the heritability (Table 1), and the effect of either a treatment or exposure (Table 2) or a disease (Table 3) on gene expression patterns. The focus of this review was on blood since it is an easily accessible tissue used by many investigators. The review identified a number of studies that have successfully uncovered new biomarkers for the pathogenesis of many diseases. Many of these studies focused on a small set of individuals and a subset of the total genes in the human genome that could have been interrogated, most likely due to the expense of microarrays. As gene expression monitoring of accessible tissues in clinical and epidemiologic studies to identify biomarkers of disease is an emerging research area, there is still much to be discovered. The expansion of concerted efforts to measure the global set of genes, in tissues other than blood, under different exposures and with a genomic set of loci, will provide a more comprehensive picture of the true underlying expression patterns.

Table 1. Example Studies Investigating Natural Variation/Heritability in Blood-based Gene Expression Patterns

Reference	Cell Type	Total Sample Size	Technology	Comments
Yan et al. 2002	lymphoblastoid cell lines	96 individuals (CEPH families)	mRNA	natural variation in gene expression
Whitney *et al.* 2003	whole blood & PBMCs	75 subjects	cDNA	natural variation in gene expression
Radich *et al.* 2004	leukocytes	32 subjects	cDNA	natural variation in gene expression
Nicholson *et al.* 2004	PBMCs	12 subjects	cDNA	natural variation in gene expression
Morley *et al.* 2004	leukocytes	122 subjects	Affymetrix	heritability of gene expression
Schadt *et al.* 2003	lymphoblastoid cell lines	56 subjects	Affymetrix	heritability of gene expression
Cheung *et al.* 2003	lymphoblastoid cell lines	90 subjects	cDNA	heritability of gene expression
Monks *et al.* 2004	lymphoblastoid cell lines	150 subjects	Affymetrix	heritability of gene expression
York *et al.* 2005	lymphoblastoid cell lines	Twin pairs	?	Twin analysis of gene expression

Key: cDNA=complementary DNA, mRNA=messenger RNA.

Table 2. Example Studies Investigating a Treatment or Exposure Blood-based Gene Expression "Fingerprint."

Reference	Cell Type	Total Sample Size	Technology	Comments
Lampe *et al.* 2004	leukocytes	85	cDNA	smoker fingerprint
Van Leeuwen *et a*l. 2005	PBMCs	7	cDNA	Smoker fingerprint
Amundson *et al*. 2004	leukocytes	8	cDNA	radiation reaction
Chon *et al.* 2004	leukocytes	21	Affymetrix	hypertension treatment study
Tang Y *et al.* 2003	whole blood	24	Affymetrix	epilepsy drug response
Jison *et al.* 2004	mononuclear cells	40	cDNA	sickle cell treatment study
Airla *et al*. 2004	PBMCs	6	cDNA	MS treatment study
Wu *et al.* 2003	lymphocytes	72	cDNA	arsenic exposure study
Whistler *et al*. 2005	PBMCs	21	cDNA	Chronic fatigue and exercise
Bittman *et al*. 2005	Whole blood	32	cDNA	Music and stress modulation

Key: cDNA=complementary DNA.

Table 3. Example Studies Seeking to Identify a Blood-based Gene Expression "Fingerprint" for Disease Phenotypes

Reference	**Cell Type**	**Total Sample Size**	**Technology**	**Comments**
Vernon *et al.* 2006	PBMCs	13	MWG-A	Symptoms of Infectious mononucleosis.
Aune *et al.* 2004	PBMCs	13	cDNA Research Genetics GF-211	Autoimmune diseases (RA, SLR, IDDM, MS)
Tang, *et al.* 2004	PBMCs	51	Affymetrix	Genetic diseases (TSC 2, NFT1, DS)
Ma *et al.* 2003	whole blood	8	cDNA	CAD study
Twine *et al.* 2003	PBMCs	45	Affymetrix	genes associated with RCC
Xu *et al.* 2004	PBMCs	?	cDNA	melanoma fingerprint
Bull *et al.* 2004	mononuclear cells	29	Affymetrix	pulmonary hypertension study
Preston *et al.* 2004	leukocytes	22	Affymetrix	IgA nephropathy
Mandel *et al.* 2004	PBMCs	36	Affymetrix	lupus study
Rus *et al.* 2003	PBMCs	33	cDNA	lupus study
Baechler *et al.* 2003	PBMCs	90	Affymetrix	lupus study
Han *et al.* 2003	PBMCs	18	cDNA	lupus study
Maas *et al.* 2002	PBMCs	36	cDNA	autoimmune study
Bennett *et al.* 2003	PBMCs	51	Affymetrix	autoimmune study
Crow *et al.* 2002	PBMCs	54	cDNA	autoimmune study
Olson *et al.* 2004	PBMCs	19	cDNA	rheumatoid arthritis onset
Tang et al. 2004	PBMCs	108	cDNA	neurofibromatosis
Satoh *et al*. 2004	PBMCs	94	cDNA	multiple sclerosis
Pellagatti *et al.* 2004	neutrophils	21	cDNA	MDS study
Vawter *et al.* 2004	lymphocytes	14	cDNA	schizophrenia study

Table 3. Continued.

Reference	**Cell Type**	**Total Sample Size**	**Technology**	**Comments**
Tsaung *et a*l. 2005	leukocytes	74	Affymetrix	schizophrenia/bipolar
Segman *et al*. 2005	PBMCs	24	Affymetrix	PTSD repeated measure study
Kalman *et al*. 2005	lymphocytes	24	cDNA	Alzheimer's disease
Tang *et al*. 2005	whole blood	129	Affymetrix	multiple neuropsychiatric diseases
Motomura *et al*. 2004	PBMCs	21	cDNA	HIV infection study
Reghunathan *et al*. 2005	PBMCs	19	Affymetrix	SARS infection study

Key: cDNA=complementary DNA, PBMCs=peripheral blood mononuclear cells, RA=rheumatoid arthritis, SLR=systemic lupus erythematosus, IDDM=insulin-dependent diabetes mellitus, MS=multiple sclerosis, TSC2=Tuberous sclerosis complex 2, NFT1=neurofibromatosis type 1, DS=Down's syndrome

The interest in the use of blood and other accessible tissues also involves the use of transformed cells and cell lines in biomarker assays. The use of transformed cells has a long history in the analysis of potential biomarkers for disease that have been tested in epidemiologic studies. No where is the more apparent than in the use of DNA repair assays and their use in predicting cancer-related disease outcomes in the population at large [37, 38]. Table 4 lists studies that have examined gene expression patterns in transformed cells from human cohorts.

Table 4. Example Studies Using Transformed Cells.

Reference	**Cell Type**	**Total Sample Size**	**Technology**	**Comments**
Yan et al. 2002	lymphoblastoid cell lines	96 individuals (CEPH families)	mRNA	natural variation in gene expression
Schadt *et al.* 2003	lymphoblastoid cell lines	56 subjects	Affymetrix	heritability of gene expression
Cheung *et al.* 2003	lymphoblastoid cell lines	90 subjects	cDNA	heritability of gene expression
Monks *et al.* 2004	lymphoblastoid cell lines	150 subjects	Affymetrix	heritability of gene expression
York *et al.* 2005	lymphoblastoid cell lines	Twin pairs	?	Twin analysis of gene expression
Vawter *et al.* 2004	lymphocytes	14 subjects	cDNA	schizophrenia study
Kalman *et al*. 2005	lymphocytes	24 subjects	cDNA	Alzheimer's disease

The Future of Association Studies

Future genetic association studies will take advantage of strategies that will both exploit available biological knowledge about the functions of genes as well as analytical methods that do not treat each variation independently. In addition, future genetic association studies will also consider phenotypic endpoints that are truly subclinical in nature, such as those derived from imaging protocols, multiplexed circulating factor (e.g., proteomic or metabolomic) analyses, or gene expression pattern analysis. To overcome problems associated with the fact that, by studying multiple variations within and across genes, as well as multiple phenotypes collected on each subject, one may have to consider the relationships between a large number of, e.g., haplotypes, multilocus genotypes, and diplotypes, and a large number of phenotypes – which could create power and interpretation issues – the procedure I have been developing may help overcome some of these issues. This dissertation considers analytical methods for association studies that exploit functional and biological data on genetic variations *and* that facilitates the grouping of individuals into more homogenous categories in ways that do not rely explicitly on reconstructing the ancestry of chromosomes or haplotypes.

I apply the proposed analysis method to publicly available gene expression data [5], and genotype data from the HapMap [3]. As emphasized, the method exploits the analysis of multiple loci simultaneously to capture their "holistic effects" by considering measures of genetic similarity, or dissimilarity (or "distances"), and a regression-like method of testing hypotheses between the genomic dissimilarity and gene expression levels (Chapter 2). I also apply the method to a whole genome association study using haplotype data obtained from the IHP (Chapter 3). I also extend the analysis technique to consider the biological coherence of gene expression biomarkers of disease across many genetic variations that may contribute to these biomarkers and their patterns in a set of individuals (Chapter 4).

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

2. (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-45

3. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437: 1299-320

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

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

6. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302

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

8. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78

9. Ioannidis JP, Gwinn M, Little J, Higgins JP, Bernstein JL, Boffetta P, Bondy M, et al. (2006) A road map for efficient and reliable human genome epidemiology. Nat Genet 38: 3-5

10. Burton PR, Tobin MD, Hopper JL (2005) Key concepts in genetic epidemiology. Lancet 366: 941-51

11. Morton NE (2006) Fifty years of genetic epidemiology, with special reference to Japan. J Hum Genet 51: 269-77

12. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4: 45-61

13. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177-82

14. Pui CH, Evans WE (2006) Treatment of acute lymphoblastic leukemia. N Engl J Med 354: 166-78

15. Low YL, Wedren S, Liu J (2006) High-throughput genomic technology in research and clinical management of breast cancer. Evolving landscape of genetic epidemiological studies. Breast Cancer Res 8: 209

16. Clementi M, Di Gianantonio E (2006) Genetic susceptibility to infectious diseases. Reprod Toxicol 21: 345-9

17. (2003) The International HapMap Project. Nature 426: 789-96

18. Salem RM, Wessel J, Schork NJ (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2: 39-66

19. Yu K, Gu CC, Province M, Xiong CJ, Rao DC (2004) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 27: 182-91

20. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231-8

21. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, Stephens JC (2003) DNA variability of human genes. Mech Ageing Dev 124: 17-25

22. Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27: 234-6

23. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-9

24. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544-8

25. Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87: 4438-41

26. (1999) Freely associating. Nat Genet 22: 1-2

27. Bird TD, Jarvik GP, Wood NW (2001) Genetic association studies: genes in search of diseases. Neurology 57: 1153-4

28. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2: 91-9

29. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598-604

30. Altshuler D, Kruglyak L, Lander E (1998) Genetic polymorphisms and disease. N Engl J Med 338: 1626

31. Tabor HK, Risch NJ, Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3: 391-7

32. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306-9

33. Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, Cucca F, Guja C, Ionescu-Tirgoviste C, Stevens H, Carr P, Nutland S, McKinney P, Shield JP, Wang W, Cordell HJ, Walker N, Todd JA, Concannon P (2002) Parameters for reliable results in genetic association studies in common disease. Nat Genet 30: 149-50

34. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG (2003) Genetic associations in large versus small studies: an empirical assessment. Lancet 361: 567-71

35. Ordovas JM (2003) Cardiovascular disease genetics: a long and winding road. Curr Opin Lipidol 14: 47-54

36. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413-7

37. Paz-Elizur T, Elinger D, Leitner-Dagan Y, Blumenstein S, Krupsky M, Berrebi A, Schechtman E, Livneh Z (2006) Development of an enzymatic DNA repair assay for molecular epidemiology studies: Distribution of OGG activity in healthy individuals. DNA Repair (Amst)

38. Fenech M (2002) Biomarkers of genetic damage for cancer epidemiology. Toxicology 181-182: 411-6

CHAPTER 2

Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis

ARTICLE

Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis

Jennifer Wessel and Nicholas J. Schork

Large-scale, multilocus genetic association studies require powerful and appropriate statistical-analysis tools that are designed to relate genotype and haplotype information to phenotypes of interest. Many analysis approaches consider relating allelic, haplotypic, or genotypic information to a trait through use of extensions of traditional analysis techniques, such as contingency-table analysis, regression methods, and analysis-of-variance techniques. In this work, we consider a complementary approach that involves the characterization and measurement of the similarity and dissimilarity of the allelic composition of a set of individuals' diploid genomes at multiple loci in the regions of interest. We describe a regression method that can be used to relate variation in the measure of genomic dissimilarity (or "distance") among a set of individuals to variation in their trait values. Weighting factors associated with functional or evolutionary conservation information of the loci can be used in the assessment of similarity. The proposed method is very flexible and is easily extended to complex multilocus-analysis settings involving covariates. In addition, the proposed method actually encompasses both single-locus and haplotype-phylogeny analysis methods, which are two of the most widely used approaches in genetic association analysis. We showcase the method with data described in the literature. Ultimately, our method is appropriate for high-dimensional genomic data and anticipates an era when cost-effective exhaustive DNA sequence data can be obtained for a large number of individuals, over and above genotype information focused on a few well-chosen loci.

Modern genetics researchers have access to an unprecedented array of technologies and resources that can be used to identify and characterize the inherited basis of disease susceptibility. For example, the availability of high-throughput sequencing and genotyping technologies, the information on the locations of ~10 million SNPs in Ensembl and related databases, and the recent release of allele-frequency and linkage disequilibrium (LD) information on >2 million SNPs by the International HapMap Project investigators have provided researchers with resources that should motivate them to pursue genetic association studies of complex, multifactorial traits and diseases, such as blood-pressure level and cancer. Unfortunately, the history of association studies that have been pursued to identify genetic variations that contribute to complex, multifactorial traits and diseases has been plagued by inconsistent results,[1] making it unclear how future large-scale association studies that are based on the use of these resources will fare. In general, the reasons for the lack of replication among association studies of complex traits and diseases are well recognized and reflect the simple fact that the influence and identification of each particular gene or environmental factor influencing these traits and diseases are often obscured or confounded by the effects of other factors. More-specific reasons for a lack of replication include differences in the choice of polymorphic sites to study, the genetic background of the population(s) sampled, the definition of the phenotype used, and the analysis methods used to assess associations.

Each of the issues plaguing association studies has been dealt with in the literature, to some degree, and new strategies are emerging that may strengthen confidence in association studies. For example, strategies for identifying appropriate polymorphisms to consider in association studies have been described by researchers involved in the International HapMap Project.[2] These strategies are based on the frequency of various alleles within and across populations, as well as the LD patterns that have emerged from analyses of them.[2] In addition, methodologies for both uncovering and accommodating population-genetic background differences and potential cryptic substructure within a specific population are being developed, in an effort to avoid false-positive and false-negative association-test results attributable to the overall genetic heterogeneity of populations.[3,4] More-sophisticated phenotyping strategies are also being developed, with an emphasis on assaying subclinical endophenotypes that may more clearly reflect pathophysiological perturbations associated with a disease and that are influenced by inherited variations.[5] The use of these phenotyping technologies is likely to accelerate the discovery of functionally relevant connections between particular genetic variations and subclinical phenotypes of all sorts.

One of the thorniest problem areas for association stud-

From the Polymorphism Research Laboratory, Department of Psychiatry (J.W.; N.J.S.), Divisions of Epidemiology (J.W.) and Biostatistics (N.J.S.), Department of Family and Preventive Medicine, The Moores Cancer Center (N.J.S.), Center for Human Genetics and Genomics (J.W.; N.J.S.), California Institute of Telecommunications and Information Technology (N.J.S.), San Diego Supercomputer Center (N.J.S.), University of California at San Diego, La Jolla; and Graduate Program in Public Health, San Diego State University (J.W.), San Diego

Received March 22, 2006; accepted for publication August 8, 2006; electronically published September 21, 2006.

Address for correspondence and reprints: Dr. Nicholas J. Schork, Department of Psychiatry, University of California at San Diego, 2062 Basic Science Building, 9500 Gilman Drive, La Jolla, CA 92093-0603. E-mail: nschork@ucsd.edu

Am. J. Hum. Genet. 2006;79:792–806. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7905-0003$15.00

ies involves relating genotype information to phenotype information in relevant statistical-analysis models. Although many analysis models and tools have been proposed in the literature, many of those tools either have been developed as extensions of traditional statistical-analysis models, such as regression models, and, as such, have inherited whatever limitations these traditional models might have (e.g., assumptions of normality), or are rooted in the exploitation of LD relationships between observed marker-locus data and unobserved trait-influencing loci. This focus on analysis methods that exploit LD is most likely the result of the current expense of genotyping individuals at a large number of loci and, therefore, the need to be economical in the choice of loci to study. We consider a complementary data-analysis strategy for genetic association studies that is based on the assessment and analysis of the similarities and differences in the allelic composition of individual genomes and the relationship of these similarities/differences to phenotypic similarities/differences. This strategy has been developed with five phenomena—related to the human genome and human physiology—in mind that, if ignored, could create problems for human association studies. We outline these five phenomena below.

First, humans are diploid and, as such, the biological effect of a gene or genes on phenotypic expression likely involves the activities and actions of both gene copies simultaneously (e.g., consider recessive-allele effects for which two copies of the allele are needed to induce a phenotype). In this light, analysis strategies that consider merely the differences in the, for example, frequency of haplotypes or alleles between individuals with and without a phenotype may be ignoring biological realities of the combined effects of the genes on the maternally and paternally derived chromosomes each individual possesses. The analysis of diplotypes, as opposed to haplotypes, however, is starting to receive attention among statistical geneticists interested in association studies.[6]

Second, it is unlikely that individual variations observed at different positions in a gene or within a group of genes function in isolation. Rather, it is more likely that the net effect of multiple variable sites in a gene or set of genes influences phenotypic expression.[7,8] Thus, there are likely subtle (if not overt) interaction effects of multiple variations within a single gene on phenotypic expression that can be observed only if one considers the influence of these variations simultaneously in an analysis.

Third, the inheritance and the evolutionary history of a set of gene variations may not be of direct relevance to the phenotypic effect of those variations. Consider, for example, the very contrived and somewhat improbable possibility that two chromosomes, each developing the same set of de novo mutations that cause a phenotype, arose in different locales at quite different times. In this situation, the assumption of a single haplotype surrounding the causal allele would be inappropriate, and methods that exploit LD patterns and common haplotypes may not work in this setting. Many analysis methods for association studies that are designed to exploit LD seek to identify and assign haplotypes and haplotype categories to individuals on the basis of the ancestry of those haplotypes (i.e., the origins of the chromosomes or haplotypes transmitted to an individual from his or her parents and the relationships of those haplotypes to a putative ancestral set of haplotypes derived from a common ancestor). However, the actual mutational or sequence profile or combination of variations in a gene and its regulatory elements are of greater relevance to association studies than are the ancestry of those variations, and the real problem lies in identifying and separating the functional variations from the neutral variations, with respect to a particular phenotype.

Fourth, each individual is likely to have his or her own unique genetic signature or combination of variations. Consider the fact that the more polymorphic sites a researcher considers in an association analysis, the greater the likelihood that each individual in the study will possess a unique pattern of variations at those sites. Thus, it is important to try to develop analysis methods for reducing the number of contrasts to be made in an association study by grouping individuals together on the basis of the common or shared variations they possess.

Fifth, studies investigating the in vitro and *in silico* functional significance of genes and genetic variations are being pursued on a large scale. These studies can shed light on polymorphic sites in the genome that are of direct relevance to a particular trait or disease.[9–11] In fact, a number of computational tools have been developed to help distinguish variations of likely functional significance on the basis of, for example, amino acid changes in an encoded protein, position in a splice site of a gene, or position in a transcription-factor binding site (TFBS).[12–15]

The association-analysis methodology that we describe attempts to address these issues by taking a more holistic multilocus diplotype view of the phenotypic effects of variations within a gene[7,16–18] and does not consider the analysis of variations as single independent factors within a gene. Our proposed association methodology considers the relationship between variation in the similarity of the allelic profile (based on alleles at polymorphic loci) among a group of individuals and additional information collected about those individuals. In this light, our methodology addresses questions such as: How much of the genomic similarity assessed with respect to variations in a particular genomic region exhibited by a group of individuals can be explained by their disease statuses and relevant ancillary information? Or, rather, is it the case that individuals with a disease or elevated values of a particular phenotype have similar genomes or genomic profiles in a region of interest that is unlikely to have arisen by chance? The method critically depends on measures of genomic similarity and dissimilarity, or "distance." These measures can be constructed in such a way as to accommodate and/or address the five aforementioned genomic and physiologic

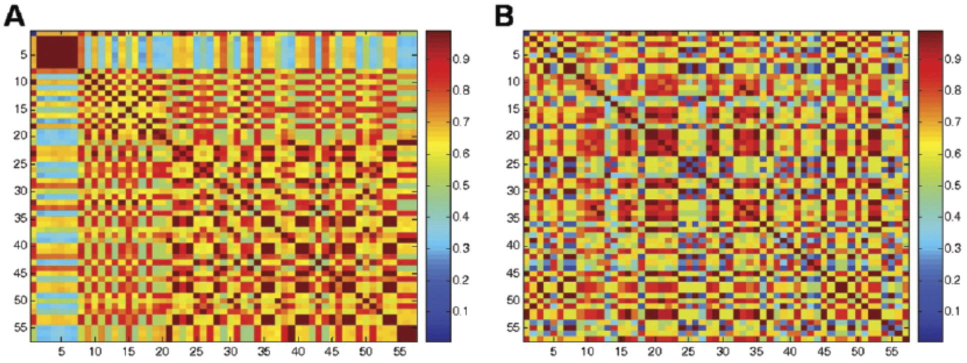


Figure 1. Heat-map representations of the similarity in the allelic profiles of 57 unrelated CEPH individuals based on variations in the *CHI3L2* gene (*A*) or the *SQSTM1* gene (*B*), with use of a standard IBS allele-sharing measure. Note that individuals have been ordered in the matrix by increasing *CHI3L2* levels. The concentration of red cells in the matrix along the diagonal in panel A suggests an association between similarity in *CHI3L2* gene composition and *CHI3L2* expression. The lack of a pattern in panel B suggests that no association exists between similarity in *SQSTM1* gene composition and *CHI3L2* expression.

phenomena not often explicitly addressed in traditional association-study data-analysis methodologies.

In describing the method, we consider the derivation of different measures of genomic dissimilarity, or distance, taking into account different features of genetic variations for each measure. We then consider the derivation of a test statistic that relates genomic dissimilarity to phenotypic end points (e.g., diagnosis, quantitative level of a phenotype, etc.). Unlike other methods, our method does not require clustering individuals into groups—which can be problematic for a number of reasons, not the least of which concerns the number of groups one should consider as present in the data. In this light, our approach is similar in orientation to the approach outlined in the derivation of the analysis of molecular variance (AMOVA) strategy of Excoffier et al.[19] However, unlike the AMOVA approach, the formulation of the model and test statistics we use are more flexible and can be used to assess multiple phenotypes, covariates, and a priori population groupings, as briefly outlined in the "Subjects and Methods" section. Our proposed method encompasses and can be used to generalize single-locus and haplotype-phylogeny analysis methods, in that one can pursue both single-locus analyses and haplotype-phylogeny analyses with the proposed procedure, as described in the "Subjects and Methods" section. In this light, our method is at least as powerful as those methods but provides possible extensions that can accommodate settings and locus effects that traditional approaches cannot. Thus, our proposed method can only improve traditional single-locus analysis of variance (ANOVA) and haplotype-phylogeny analysis methods. We describe data sets used to showcase the proposed techniques and the results of relevant analyses of these data sets. We end with a discussion and considerations of areas for future research.

Subjects and Methods

Measures of Genomic Similarity

There are a number of strategies for characterizing the similarity of individuals with respect to the variations they possess, both within and across different genes. We describe seven example methods for assessing the genomic similarity between two individuals on the basis of genotype data. Some of these methods have been designed to accommodate weighting schemes for various factors, such as allele frequency or locus functional significance. In addition, it is possible that combinations of the approaches could be pursued (e.g., weighting by both frequency and function). Weighted similarity measures have been extensively studied in cluster-analysis contexts and so are appropriate to consider in other contexts.[20] Once a similarity measure has been chosen, it can be evaluated for all pairs of N individuals in a sample, to construct an $N \times N$ similarity matrix, where element i,j of that matrix contains the similarity value for individuals i and j ($i,j = 1,\ldots,N$). We note that other groups have considered different measures of genomic similarity that may be of value[21] (see, e.g., the works by Müller et al.[22] and by Sielinski[23]). We also note that similarity matrices admit intuitive graphical representations in the form of heat maps and trees,[24–26] which makes our proposed analysis procedure intuitively appealing, as described below (see also figs. 1 and 2).

Similarity based on identity-by-state (IBS) allele sharing.—The fraction of alleles that any two individuals share purely by state (e.g., the two individuals possess the same allele or variant at a locus) can be calculated easily enough. Since humans have two copies of each position on the genome, it is simple to determine how many alleles (0, 1, or 2) a pair of individuals shares. By dividing by twice the number of loci or positions studied, one can obtain an estimate of the fraction of alleles shared IBS by those individuals. Pairwise similarities derived in this manner have been used

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 2. Tree representation of the similarity in the allelic profiles of 57 unrelated CEPH individuals. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

to construct a matrix for cluster (and related) analyses, to address population genetic research questions.[27,28] The IBS-sharing similarity, $S_{i,j}$, can be calculated for individuals i and j $(i,j, = 1, \dots ,N)$ with the formula

$$S_{i,j}^{\mathrm{IBS}} = \frac{\sum_{l=1}^{L} s_{i,j}^{l}(g_i^l,g_j^l)}{2L} \ , \qquad (1)$$

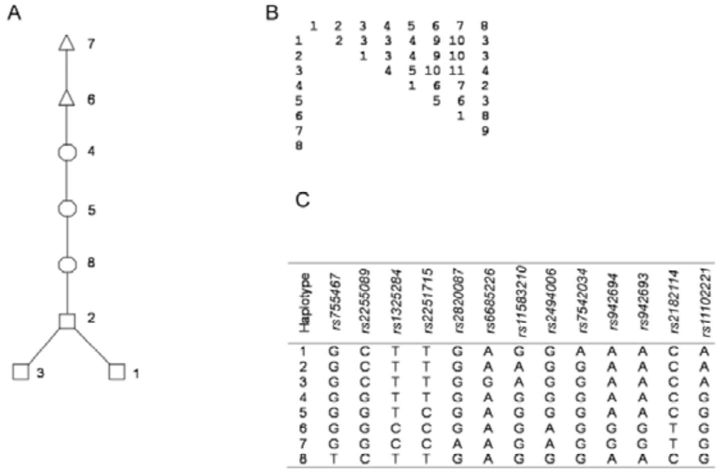
where L is the number of loci considered in the calculation; g_i^l and g_j^l are the genotypes of individuals i and j, respectively, at the lth locus $(l = 1, \dots ,L)$; and $s_{i,j}^{l}(g_i^l,g_j^l)$ is a function mapping the genotype information, for individuals i and j at locus l, to a particular numeric value and, for our purposes, has a value of 0.0 if individuals i and j are homozygous for different SNP alleles (e.g., $g_i^l = AA$ and $g_j^l = TT$), a value of 1.0 if they share one allele (e.g., $g_i^l = AA$ and $g_j^l = AT$), and a value of 2.0 if they share both alleles (e.g., $g_i^l = AA$ and $g_j^l = AA$)—note that we are assuming, throughout, that interest is in SNP loci with two alleles, as opposed to microsatellite markers and other forms of genetic variation, although the proposed method can be easily extended to cover situations in which those forms of variation are examined.

Similarity based on weighting by allele frequency.—Allele-frequency information can be included in the construction of the measure of similarity. The intuition behind the accommodation of allele frequency is that individuals who share rare alleles may have more-similar genomes than do individuals who share common alleles (i.e., since many people will have common alleles, individuals possessing them are not easily distinguished from others). Lynch and Ritland devised a method (hereafter called "the LR method") for assessing genomic similarity, on the basis of genotype data, that accounts for allele frequency and has been shown to have some favorable properties for identifying population subgroups.[21,29,30] With notation derived from equation (1), weighted similarity measures can be computed easily as

$$S_{i,j}^{w} = \frac{\sum_{l=1}^{L} w^l s_{i,j}^{l}(g_i^l,g_j^l)}{\sum_{l=1}^{L} w^l} \ , \qquad (2)$$

where w^l is a positive number reflecting the weight assigned to locus l.

IBS allele sharing, with weighting for functionality of variations.—One can accommodate knowledge of the "functional significance" of variations in a measure of genomic similarity by giving greater weight in the sharing measure to loci harboring functional variations. These weights must be determined a priori and can be based on, for example, the results of cellular in vitro assays investigating the influence of variations on gene expression or protein-binding potential. As an example, consider a situation in which in vitro functional-analysis assays suggest that variations at two polymorphic sites in the promoter region of a gene resulted in a 1.5-fold and a 2.0-fold increase in expression levels and that a variation at another polymorphic site resulted in a protein amino acid change that causes a binding site in that protein to induce a 2.0-fold increase in activity of the protein. In this hypothetical situation, one could assign weights of 1.5, 2.0, and 2.0, respec-



	1	2	3	4	5	6	7	8
1		2	3	3	4	9	10	3
2			1	3	4	9	10	3
3				4	5	10	11	4
4					1	6	7	2
5						5	6	3
6							1	8
7								9
8								

Haplotype	*rs7555467*	*rs2255089*	*rs13252284*	*rs2251715*	*rs2820087*	*rs6685226*	*rs11583210*	*rs2494006*	*rs7542034*	*rs942694*	*rs942693*	*rs2182114*	*rs11102221*
1	G	C	T	T	G	A	G	G	A	A	A	C	A
2	G	C	T	T	G	A	A	G	G	A	A	C	A
3	G	C	T	T	G	G	A	G	G	A	A	C	A
4	G	G	T	T	G	A	G	G	G	A	A	C	G
5	G	G	T	C	G	A	G	G	G	A	A	C	G
6	G	G	C	C	G	A	G	A	G	G	G	T	G
7	G	G	C	C	A	A	G	A	G	G	G	T	G
8	T	C	T	T	G	A	G	G	G	A	A	C	G

Figure 3. *A,* Tree representation of the phylogenetic relationships of haplotypes derived from the *CHI3L2* genotype data for the 57 unrelated CEPH individuals, with use of the method of Seltman et al.[35,38] *B,* The distance matrix used to construct the phylogenetic tree, with the numbers on the rows and columns identifying the different haplotypes. Note that the haplotypes are identified with numbers assigned arbitrarily (*C*). Haplotypes that are phenotypically similar are denoted by their corresponding symbols.

tively, to the loci harboring these variations, in the construction of the similarity measure.

Similarity based on weighting by nucleotide conservation across species.—In the absence of data on the potential functional effect of variations, one could consider a criterion for weighting loci in a genomic similarity measure that is based on conservation of nucleotides across species. It has been argued that nucleotides that are conserved throughout evolution are more likely to be of functional significance, since changes at those positions may have undergone negative selection; see, for example, the works of Shah et al.,[31] Frazer et al.,[32] and Brudno et al.[33] Thus, one could weight genomic positions used in a genotype-based similarity measure by the degree of evolutionary conservation at those positions.

Similarity based on single-locus–analysis results.—We consider the use of single-locus–analysis results in the construction of a multilocus similarity measure. Single-locus analyses can be pursued before the construction of a similarity measure or could be based on analyses performed previously with another data set. We consider the use of the negative log of the P value associated with single-locus–analysis test statistics as weights in the construction of a similarity measure using, for example, equation (2) or equation (3).

Unweighted and weighted haplotype-pair similarity.—By phasing individuals (i.e., assigning them haplotypes that reflect variations they inherited on their maternally and paternally derived chromosomes), one can assess the similarity of two individuals' chromosome pairs. A relevant similarity measure would depend critically on how one pairs (or matches) the chromosomes between the two individuals, since the similarity could be very different for the two possible pairings. A better measure would involve computing the similarity with the assumption of both pairings and then taking the maximum measure that results from these two pairings as the measure of similarity. Consider, for example, the simple situation in which individual a has haplotypes ha1 = 0-0-0-0 and ha2 = 1-1-1-1 and individual b has haplotypes hb1 = 1-1-1-1 and hb2 = 0-0-0-0. Then, to assess haplotype similarity, if one pairs ha1 with hb1 and pairs ha2 with hb2, the individuals would have completely different genomes (i.e., have maximal distance, or zero similarity). However, if one pairs ha1 with hb2 and pairs ha2 with hb1, then the individuals have identical genomes. We believe that use of the pairing that maximizes the similarity is appropriate, and that was the motivation for the measure reflected in equation (3). Haplotype-based sharing can easily accommodate weighting schemes based on, for example, conservation or functionality, in which some loci have been weighted because of their putative functionality. Although slightly more complicated than the genotype similarity–based measures, haplotype pair–similarity measures can be computed as

$$S^h_{i,j} = \max\left\{\sum_{l=1}^{L} w^l[s^l_{i,j}(h^l_{i,1},h^l_{j,1}) + s^l_{i,j}(h^l_{i,2},h^l_{j,2})],\right.$$
$$\left.\sum_{l=1}^{L} w^l[s^l_{i,j}(h^l_{i,1},h^l_{j,2}) + s^l_{i,j}(h^l_{i,2},h^l_{j,1})]\right\}$$
$$\times\left(\sum_{l=1}^{L} w^l\right)^{-1}, \qquad (3)$$

where the similarity function considers the alleles on specific haplotypes possessed by individuals i and j and would assign a numerical value of 0.0 if the individuals did not have the same allele on those haplotypes and 1.0 if they did (note that, in eq. [3], $h^l_{i,1}$ refers to individual i's allele at position l of his or her chromosome designated as 1, as opposed to 2). In addition, the fact that one could pair the first haplotype (arbitrarily defined) possessed by individual i with either the first or the second haplotype possessed by individual j is accommodated in the calculation by use of the maximum of these two pairings to define the similarity.

Similarity based on ancestry.—There are many association-analysis methods that consider similarity in the phylogenetic connections or ancestry of haplotypes.[34–37] For example, the programs eHAP,[35,38] HAP,[39] Arlequin,[40] and GeneTree[41] produce phylogenies of chromosomes on the basis of genotype data. The phylogenies produced by these programs can then be used to group individuals into smaller subgroups that can be used to contrast phenotypic features. We consider this approach as an alternative to those that are based on, for example, functional-variation similarity, although recent studies have suggested that grouping haplotypes on the basis of phylogeny and then contrasting the resulting groupings for phenotypic differences does not substantially increase power to detect an effect (see, e.g., the work of Humphreys and Iles[42] and Bardel et al.[43]). However, we note that one can exploit ancestral relationships between haplotypes to derive a similarity measure. Essentially, from a phylogenetic tree, one can determine the distance between haplotypes (e.g., on the basis of the number of mutations, recombinations, gene conversions, or transitions that must have occurred to derive one haplotype from another ancestral haplotype) (see fig. 3). With this information, one can pair the haplotypes that two individuals possess and can compute the phylogenetic distance between those haplotypes. Since this pairing can occur in two ways, we take the pairing that produces the minimum distance between the two individuals as reflecting the similarity between them, as was done for the haplotyping pairing–similarity measures discussed above.

Regression-Based Distance Matrix Analysis

Once one has computed a similarity matrix, that matrix can be subjected to a regression analysis that tests hypotheses about whether variation in the level of similarity exhibited by pairs of individuals reflected in that matrix can be explained by other features those individuals possess (e.g., whether they possess a certain phenotype or have higher or lower values of a particular quantitative phenotype). To describe the regression model, we consider an analysis involving a gene or genomic region that harbors L different polymorphic loci. We also assume that each of N individuals or study subjects has been genotyped at these L loci. We assume also that M phenotypic variables have been collected on the N subjects. These phenotypic variables could include information about the presence or absence of a disease end point (e.g., coded using dummy variables, such as 0 assigned to individuals without the disease and 1 assigned to individuals with the disease); disease-associated quantitative variables, such as blood pressure and cholesterol level; and important covariates, such as age, sex. smoking status, etc. We assume that interest is in relating the disease end points or quantitative variables to the genomic profiles of the individuals, as captured by the genotypic information collected about them.

Construct an N × N similarity matrix with, for example, one of the measures described in the "Measures of Genomic Similarity" section. Transform the matrix into a dissimilarity, or "distance," matrix by, for example, subtracting the components of the matrix from 1.0 if the IBS measure is used or by subtracting them from 1.0 after each component in the matrix is divided by the theoretical or empirical maximum of the similarity measure, to scale the entries to lie between 0 and 1. We note that the proposed regression procedure does not require that the distance matrix have metric properties.[44] Let this distance matrix and its elements be denoted by $D = d_{ij}$ $(i,j = 1, \ldots ,N)$, for the N subjects. The possibility that $N \ll L$ will not pose problems in the proposed regression-analysis setting. Let X be an $N \times M$ matrix harboring information on the M phenotypic variables that will be modeled as predictor or regressor variables whose relationships to the values in the genomic similarity matrix are of interest. Compute the standard projection matrix, $H = X(X'X)^{-1}X'$, typically used to estimate coefficients relating the predictor variables to outcome variables in multiple-regression contexts. Next, compute the matrix

$$A = (a_{ij}) = (-\frac{1}{2}d_{ij}^{2}) \ ,$$

center this matrix with use of the transformation discussed by Gower,[45] and denote this "matrix G" as

$$G = \left(I - \frac{1}{n}11'\right)A\left(I - \frac{1}{n}11'\right) \ .$$

An F statistic can be constructed to test the hypothesis that the M regressor variables have neither relationship to variation in the genomic distance nor dissimilarity of the N subjects reflected in the $N \times N$ distance/dissimilarity matrix as done by McArdle and Anderson[46]:

$$F = \frac{tr(HGH)}{tr[(I-H)G(I-H)]} \ . \qquad (4)$$

If the Euclidean distance is used to construct the distance matrix on a single quantitative variable (i.e., as in a univariate analysis of that variable) and appropriate numerator and denominator degrees of freedom are accommodated in the test statistics, then the F statistic in equation (4) is equivalent to the standard ANOVA F statistic.[46] The distributional properties of the F statistic are complicated for alternative distance measures computed for more than one variable, especially if those variables are discrete, as in genotype data. However, permutation tests can then be used to assess statistical significance of the pseudo–F statistic.[44,46–50] The M regressor variables can be tested individually or in a stepwise manner.

Graphic Display of Similarity Matrices

Similarity matrices of the type we describe can be represented graphically in a number of ways that can facilitate interpretation. We consider heat-map and coded-tree (or dendogram) representations.[24–26] Heat maps simply color code the elements of a similarity matrix, such that higher similarity values are represented as "hotter," or redder colors, and lower similarity values are represented as "colder," or bluer colors. If the matrix is ordered such that individuals with similar phenotype values are next to each other, then neighboring cells along the diagonal of the matrix (representing individuals with similar phenotype values) will present patches of red, indicating a relationship between the phenotype values and similarity (fig. 1*A* and 1*B*). Trees are constructed such that individuals with greater genomic similarity are placed next to each other (i.e., they are represented as adjacent branches of the tree). Less similar individuals are represented as branches some distance away from each other. By color coding the individual branches on the basis of the phenotype values of the individuals they represent, one can see if there are patches of a certain color on neighboring branches, which would indicate that phenotype values cluster along with genetic similarity (fig. 2)

The CEPH Family Gene-Expression Data as an Example Data Set

To showcase the proposed method relative to other methods, we considered an analysis involving gene-expression and SNP data collected on 57 unrelated CEPH individuals. These individuals were chosen by HapMap researchers for massive, genomewide genotyping studies[2] and were also used to assess gene-expression patterns obtained from immortalized lymphocytes[51] (Gene Expression Omnibus accession number GSE2552). Our analysis excluded individual NA06993 in the gene-expression studies, because detailed analysis of HapMap data suggested that the sample associated with this person is likely to have derived from an unreported relative. We also added data associated with individual NA12056, since gene-expression data for this individual is now available. We focused on the analysis of variations genotyped on the CEPH individuals in the *CHI3L2* gene (MIM 601526), since Cheung et al.[51] found very compelling evidence of association and linkage to this gene for the expression levels of the *CHI3L2* gene, reflecting likely *cis*-acting sequence variations influencing expression of the encoded protein. We downloaded, from the HapMap database, data on 43 SNPs in the *CHI3L2* gene that were genotyped on the 57 CEPH individuals with *CHI3L2* gene–expression values (chromosome 1: 111069007–111084786; Ensembl position 111482322–111498101). We derived the positions of the SNPs from the latest version of the human physical map provided in Ensembl. We note that these positions disagree slightly with those reported by Cheung et al.[51]

Haplotyping and Basic Analysis of the Expression Data

Haplotypes and diplotypes (i.e., the pairs of haplotypes each individual possesses) were inferred using HAP.[39] Repeated, multiple gene–expression values collected for each of the CEPH individuals were averaged, when available. We considered use of $\log_2$-transformed expression levels because of skewness in the expression values. We assessed the association between the SNPs in the *CHI3L2* gene and *CHI3L2* gene–expression values, using regression analysis of each, coded as 0, 1, or 2, depending on how many minor alleles each individual possessed at a SNP locus. We also tested for haplotype associations, using the haplotype-phylogeny–analysis methods described by Seltman et al.[35,38] We then applied the proposed analysis method using different similarity measures. We also included analyses that considered each locus in isolation, using the proposed similarity-regression procedure—that is, we constructed the similarity matrix using genotype information for each locus independently. To correct for multiple testing in the single-locus analyses, we used the method developed by Nyholt,[52] to determine the "effective" number of independent SNPs from the total of 26 that we studied. The effective number of independent SNPs was found to be $\sim$14. We then used a Sidak-corrected P value to declare significance at the nominal level of $P < .05$. For the distance-based regression analysis, we used 100,000 permutations of the data to assess the probability of a type I error. We considered sex as a covariate in the analysis methods used.

Assessing SNP Functional Significance for Similarity-Measure Weighting

For the proposed similarity measure exploiting functional information on the SNPs, we considered the use of a number of resources, including results of in vitro studies, (e.g., promoter-reporter cell transfection and/or model species analyses), *in silico* (computational) structure and sequence analysis, and sequence-conservation analyses. We considered SNPs in all the genetic regions that were available—for example, putative functionally relevant SNPs, such as coding (synonymous and nonsynonymous) and noncoding (exonic splicing enhancers [ESEs] and TFBSs, in both the $5'$ and $3'$ UTRs), as well as likely neutral variations, such as SNPs in functionally obscure intronic sites. We used available Web-based tools and programs to assess functionality (see table 1, which describes the analysis tools and references for our assessment of functionality and conservation). To complement the information we obtained from individual Web sites, we used PupasView, a Web site that gives comprehensive functional information from many individual programs and databases. To assess evidence of evolutionary sequence conservation at the site harboring each SNP, we leveraged data from multiple species. Genomic regions that show evidence of multiple-species sequence conservation at the nucleotide level are more likely to have undergone selective pressures and, hence, are likely to be of functional significance.[53] Use of multiple-species genomes in comparisons with the human sequence (as the reference genome) has the advantage of providing stricter criteria, which minimizes false-positive conservation results with any one species and improves the ability to classify elements as actively conserved because of functional consequences rather than shared ancestry. For weighting based on sequence conservation, sites had to be identified as conserved in two or more species to be given a greater weight; when sites were found conserved across more than two species, more weight was given. In using functional information to weight the SNPs in a similarity measure, we intentionally kept our weighting scheme simple so as not emphasize the absolute value of the weight; rather, the weighting was relative to each SNP. We felt that the most weight should be given to SNPs characterized as functional by in vitro methods (e.g., SNP *rs*755467 = 2.0) (see table 2). If multiple *in silico* methods identified a SNP as having plausible functional consequence, then more weight was given to that SNP than to those for which only one method suggested functionality. In addition, because of imprecision in the computational identification of regulatory binding motifs, we required agreement among programs used for identifying a sequence as being in a TFBS, a UTR, or an ESE, to reduce false-positive results.

Results

Polymorphic Variation in CHI3L2

In *CHI3L2,* 11 SNPs were monomorphic, and 6 SNPs were excluded because of low minor-allele frequency (<2.0%) (data not shown). Five SNPs were not in Hardy-Weinberg equilibrium (HWE) ($P < .05$) (table 2). Because of the thorough quality assessment and control of the data by the HapMap researchers, we assumed that those SNPs not exhibiting HWE were not an artifact of genotyping; therefore, we opted to keep them in the analyses but also conducted analyses that excluded those SNPs. Four SNPs were tagging SNPs (tSNPs), on the basis of HapMap analyses. Five SNPs in the *CHI3L2* gene were coding SNPs, as reported to dbSNP, one of which was genotyped by the HapMap researchers. The other SNPs were in noncoding or regulatory regions. The majority of the SNPs were in strong LD (average $D' = 0.97$), with the exception of SNP Thr313Thr, which showed weaker LD ($D' = 0.02$–0.80) with nine other SNPs (data not shown but easily visualized on the HapMap site).

SNP Functionality and Sequence Conservation Assessment Results

In Vitro and In Silico *Analysis of* CHI3L2 *Variation.*—Promoter activity of luciferase reporter assays containing the *rs755467* SNP was twofold higher (T→G allele) than that of constructs not containing the *rs755467* SNP. This increase in promoter activity was because of stronger binding of RNA polymerase II.[51] Four SNPs were identified, through *in silico* methods, as being in potential ESE sites (table 2). Although analyses involving the Web site PupasView found results

Table 1. Resources for Identifying or Predicting Function and Conservation in *CHI3L2*

Reference	Function	Comment
Cheung et al.[51]	In vitro	Twofold increased binding of RNA polymerase II (T→G allele)
PupasView	Comprehensive	
SIFT	Nonsynonomous SNPs	Sorting Intolerant from Tolerant
PolyPhen	Nonsynonomous SNPs	
ESEfinder	ESEs	
RESCUE-ESE	ESEs	
Gene Regulation	TFBSs	For P-Match
Vista Tools	TFBSs	For rVISTA
UTRScan	UTR functional elements	
ITB Blast	UTR functional elements	For BigBlast
VISTA Genome Browser	Conservation	
PipMaker and MultiPipMaker	Conservation	For PipMaker

Table 2. Characteristics of *CHI3L2* SNPs, Functional Consequences, and Conservation

SNP	Ensembl Position	Location	Minor-Allele Frequency	HWE *P*	tSNP	ESEfinder[a]	RESCUE-ESE[a]	Functional Weight	Frog	Chicken	Mouse	Cow	Opossum	Dog	Fugu	Chimpanzee[b]	Conservation Weight
rs755467[c]	111482465	Intron 1	.28	.739	N		Y (1)	2				62				99	1.15
rs2147790	111482633	Intron 1	.16	.082	N			1				62				99	1.15
rs2255089	111485610	Intron 3	.46	.253	N	Y (1)		1	34			63	20		18	97	1.30
rs2274232	111485642	Intron 3	.11	.003	N		Y (1)	1	34		63	63	20		18	97	1.45
rs2147789	111485872	Intron 3	.44	.025	Y		Y (3)	1	34			63	29			100	1.25
rs2182115	111486179	Intron 4	.09	.011	N	Y (2)		1	26		33	64	28			98	1.30
rs1325284	111487834	Intron 4	.33	1.000	N	Y (1)	Y (3)	1.5				70	28			98	1.65
rs2251715	111490229	Intron 5	.44	.274	N			1				72	35			98	1.50
rs961364[d]	111490510	Intron 6	.26	.613	N	Y (1)	Y (1)	1.75	19		19	71	42	31		99	1.70
rs2764543	111491140	Intron 7	.31	.771	N			1				68				98	1.15
rs7366568	111491806	Intron 7	.25	.020	N	Y (4)		1			24					98	1.10
rs2820087	111492376	Intron 7	.28	.975	N		Y (2)	1	37							98	1.10
rs6685226	111492646	Intron 7	.17	.142	N	Y (1)		1	51	70						98	1.35
rs11583210	111493928	Intron 8	.25	.075	Y			1	29		30	73	28	25		98	1.60
rs12032329	111494414	Intron 8	.12	.003	N		Y (2)	1	29		30	71				98	1.55
rs2477578	111495313	Intron 8	.33	1.000	N	Y (1)		1	29			71			13	98	1.55
rs2494006	111495483	Intron 8	.28	.956	N	Y (2)		1	29			71				98	1.50
rs7542034[e]	111496023	Thr313Thr	.02	.888	N	Y (3)		1.35	60		62	88	79		61	97	1.75
rs942694	111496180	Intron 9	.33	1.000	N			1	30		16	52	25			98	1.30
rs942693	111496200	Intron 9	.33	1.000	N	Y (2)		1	30		16	52	25			98	1.30
rs2182114	111496269	Intron 9	.33	1.000	N	Y (1)		1	30		16	52				98	1.40
rs5003369	111496447	Intron 9	.33	1.000	Y	Y (1)		1	30			52				98	1.20
rs11102221	111496858	Intron 9	.26	.161	N		Y (1)	1	30			71	38			98	1.55
rs3934922	111497436	Intron 10	.30	.556	Y	Y (3)	Y (2)	1.25	34					72		99	1.50
rs3934923	111497509	Intron 10	.33	1.000	N	Y (1)		1	34					72		99	1.65
rs8535	111497971	Exon 10	.28	.739	N	Y (1)		1	32					64		100	1.20

NOTE.—*In silico* results from UTR, TFBS, and nonsynonomous SNPs are omitted, since no functional SNPs were identified. Twelve species were considered in VISTA Genome Browser, five with high conservation (>70% identity). Fifteen species were considered in PipMaker; *Drosophila* results from PipMaker were omitted (62% conservation at *rs7542034*).

[a] The number of sequence motifs identified by the ESEfinder and RESCUE-ESE programs is shown in parentheses.

[b] Chimpanzee results are combined from VISTA Genome Browser and PipMaker.

[c] In vitro results showed 2 times greater binding by RNA polymerase II to the T allele, compared with the G allele.

[d] PupasView finding was a triplex, a possible regulatory element.

[e] PupasView finding was an ESE (3).

similar to the analysis results based on other Web-based tools, it also uniquely identified a triplex, a possible regulatory element. UTRScan and BIGBlast did not identify any SNPs in UTR-binding motifs, and, similarly, P-Match (Gene Regulation) and rVISTA (Vista Tools) did not identify any TFBSs (results not shown). There were no nonsynonomous SNPs for evaluation.

Sequence Conservation Analysis of CHI3L2 *Variation.*—VISTA browser2 (VISTA Genome Browser) was used to compare the sequence of *CHI3L2* (chromosome 1: 111482202–111498000) from human (reference genome, May 2004 build, except that the chimpanzee sequence was compared with the July 2003 build) with 10 other available species. Overall, the amount of highly conserved sequence decreased with increasing phylogenetic distance. Chimpanzee, cow, dog, opossum, and fugu exhibited some highly conserved regions with humans; chimpanzee had the largest number of conserved sequences, and fugu had the least (>70% identity) (note that fugu has conserved sequence but does not contain any of the SNPs investigated in the present study). When we reduced to 5% the allowable percentage of identity for analysis, there was moderate conservation with mouse, frog, and chicken, and there was no conservation with rat or zebra fish. PipMaker and MultiPipmaker were also used for pairwise and multiple-species comparisons and to extend to different species not available at the VISTA Genome Browser Web site. PipMaker and VISTA Genome Browser pairwise comparisons gave slightly different results because of the different algorithms used (local vs. global homology, respectively). Only VISTA Genome Browser results are shown. Nine SNPs were in highly conserved regions in two or more species, whereas some conservation was found with other species (table 2). For comparisons involving a species more evolutionarily distant from humans, we studied sequence from *Drosophila melanogaster.* We found that only one SNP, Thr313Thr—the SNP that was most consistently conserved across multiple species—was in a conserved region, which suggests it might be a functionally important region for this gene. *Anopheles gambiae* was compared with human, and no conserved sequences were identified (data not shown). Comparison across multiple species, which can identify conserved regions possibly under selection, revealed SNPs with sequences conserved across multiple species and with more-distant species (*rs7542034* and *rs2255089*).

To assign weights that were based on this information, we used a minimum identity of 70% to locate SNPs that were in highly conserved regions. Then, we lowered the minimum threshold to 5%, to identify regions of moderate conservation (40%–69%) or low conservation (10%–39%). Under the assumption that the in vitro results give the most-compelling results, our conservation weighting was scaled from 1 to 1.75, where 1 is no conservation across multiple, pairwise species comparisons and 1.75 is the most conserved region (with high conservation and the most species) and represents a value less than those of the in vivo results. We categorized conservation levels as high, moderate, or low and used this scheme to assign weights to the loci. We recognize that our scheme for assigning weights may seem arbitrary, but we chose to expose the use of weights that are based on different criteria and not necessarily to focus on the optimal manner in which weights can be assigned.

Single Locus–Analysis Results

Of the 26 SNPs, 14 were significantly (i.e., $P \leqslant .001$) associated with *CHI3L2* $\log_2$-transformed expression levels after correction for multiple tests (note that none of these analyzed SNPs deviated from HWE) (table 3 [columns 3–6]). In addition, two of the four SNPs identified as tSNPs, according to the HapMap Web site, were significantly associated with *CHI3L2* levels. We also include in table 3 the results of the single-locus analyses with use of the proposed similarity-analysis approach (columns 7–10), and, as can be seen, the single-locus results with the proposed procedure correspond well with results obtained from the traditional regression-based single-locus analysis. We note that the Spearman rank correlation between the *P* values obtained from these two analyses was 0.862 ($P < .0001$). Thus, our proposed procedure can be used to conduct single-locus analyses.

Haplotype Associations and Haplotype-Phylogeny Results

We analyzed the data, using eHAP, a program that infers haplotypes and implements evolutionary-based association analyses (fig. 3*A*). eHAP constructs a cladogram that is based on the method described by Templeton et al.[34] and then performs sequential association testing between "nearby" haplotype clades, collapsing them and grouping them together if no trait differences are found between the two haplotype groups, given the others. Because of algorithmic limitations, redundant SNPs ($D' = 1.0$) were deleted by choosing the least informative (no functionality) SNPs to represent a group. Eight common haplotypes consisting of 13 SNPs were identified with frequencies of 1.5%–30.3% (fig. 3*C*). The final grouping that showed the maximum phenotypic difference in *CHI3L2*-expression levels consisted of haplotypes 6 and 7 versus haplotypes 1–5 and 8 ($P = .0009$). This latter grouping contains the minor alleles of the two most functionally important SNPs (*rs755467* and *rs7542034*) that have in vitro and strong conservation evidence. The distance matrix calculated as part of the analysis implemented in the eHAP program (fig. 3*B*) was also used to construct a measure of similarity between individuals.

Similarity Regression-Analysis Results

Analysis of multiple loci in CHI3L2.—Significant associations between the values in the genomic-similarity matrix and gene-expression levels were found with each of the measures of genetic similarity (table 4) ($P < .001$). Most no-

Table 3. Individual SNP Associations with *CHI3L2* Expression Levels

		Traditional Regression Analysis					Similarity Regression Analysis		
SNP	Location	*F*	Exact *P*	Corrected *P*[a]	Variation (%)	$(-)\log_{10} P$	IBS *F*	IBS *P*	Variation (%)
rs755467	Intron 1	15.57	4.58×10^{-6}	.00006	37	4.19	3.71	.00001	36
rs2147790	Intron 1	.28	7.53×10^{-1}	1.00000	1	.00	.17	.68502	00
rs2255089	Intron 3	3.68	3.16×10^{-2}	.36208	12	.44	6.49	.01408	11
rs2274232	Intron 3	1.01	3.70×10^{-1}	.99845	4	.00	2.01	.16097	4
rs2147789	Intron 3	5.44	7.70×10^{-3}	.10257	19	.99	8.66	.00477	16
rs2182115	Intron 4	2.23	1.17×10^{-1}	.82511	8	.08	4.50	.03784	8
rs1325284	Intron 4	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs2251715	Intron 5	3.28	4.52×10^{-2}	.47667	11	.32	6.38	.01496	10
rs961364	Intron 6	15.37	6.94×10^{-6}	.00010	39	4.01	31.38	.00001	39
rs2764543	Intron 7	12.99	2.57×10^{-5}	.00036	33	3.44	15.59	.00025	22
rs7366568	Intron 7	1.35	2.51×10^{-1}	.98254	3	.01	1.35	.25222	3
rs2820087	Intron 7	1.86	1.00×10^{-4}	.00140	31	2.85	15.23	.00023	23
rs6685226	Intron 7	.41	6.69×10^{-1}	1.00000	2	.00	.42	.51809	1
rs11583210	Intron 8	1.44	2.45×10^{-1}	.98048	5	.01	2.66	.10922	5
rs12032329	Intron 8	.89	4.15×10^{-1}	.99945	3	.00	1.79	.18856	3
rs2477578	Intron 8	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs2494006	Intron 8	15.26	7.82×10^{-6}	.00011	39	3.96	14.12	.00046	23
rs7542034	Thr313Thr	.00	9.66×10^{-1}	1.00000	0	.00	.00	.96759	0
rs942694	Intron 9	16.99	2.22×10^{-6}	.00003	40	4.51	2.08	.00006	28
rs942693	Intron 9	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs2182114	Intron 9	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs5003369	Intron 9	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs11102221	Intron 9	1.39	2.57×10^{-1}	.98432	5	.01	2.56	.11535	4
rs3934922	Intron 10	13.45	1.82×10^{-5}	.00025	33	3.59	26.96	.00001	33
rs3934923	Intron 10	16.85	2.06×10^{-6}	.00003	38	4.54	19.97	.00008	27
rs8535	Exon 10	15.57	4.58×10^{-6}	.00006	37	4.19	3.71	.00001	36

NOTE.—Association analysis with averaged, $\log_2$ gene-expression levels.

[a] *P* value corrected for multiple tests, with use of SNPSpD[52] to find the effective number of SNPs and with use of Sidak's method to find the experimentwise error rate with individual SNPs.

table were the analyses involving weighted associations in which the weighting was based on functionality ($P = .00006$), as well as weighting by association-strength ($P = .00001$) allele sharing. The LR allele-frequency weighted measure gave the least significant association. Similar results were found for allele sharing and haplotype sharing, suggesting that phasing might not aid in detecting associations at this locus. Similar results were found with use of all 26 SNPs versus the subset that excluded the 5 SNPS not in HWE, although slightly stronger associations were found with the 21 SNPs in HWE, which suggests that HWE could influence correct calculation of genetic similarity or that the use of too many SNPs could dilute the association effect. The five SNPs that depart from HWE appeared to have no or little functional consequence. When we restricted analyses to the four tSNPs and calculated genetic similarity by allele sharing, we found similarly significant results.

Analysis of nonassociated genes and phenotypes.—To show that our method is not too liberal in identification of associations, we studied two genes, *SQSTM1* and *GSTM2,* whose expression levels were not correlated with *CHI3L2* expression or *CHI3L2* polymorphisms. *GSTM2* was reported as having significant *cis*-acting SNPs that influenced *GSTM2*-expression levels and is near *CHI3L2*; *SQSTM1* did not have *cis*- or *trans*-acting SNPs that affect its expression levels. Four SNPs in *GSTM2* and 12 SNPs in *SQSTM1* were downloaded from the HapMap Web site. Two and six SNPs were excluded from the analyses of *GSTM2* and *SQSTM1,* respectively, because they were monomorphic. For individual SNP analyses with SNPs in *SQSTM1* and $\log_2$-transformed *SQSTM1*-expression levels, no SNPs were significant ($P > .15$). When analyses with genetic similarity by allele sharing or allele-frequency weighted sharing were performed, associations did not meet the threshold of significance ($P = .3818$ and $.3848$, respectively) (table 4). With SNPs in *GSTM2* and *CHI3L2*-expression levels, no significant associations were found with individual SNPs, allele-sharing similarity, or allele-frequency weighted allele sharing similarity ($P > .38$) (table 4).

Assessing Association Signal Strength and Detecting Interactions

We pursued a few additional studies, to assess the merits of the proposed analysis procedure. First, we considered the contribution of each *CHI3L2* SNP to the association strength by removing each SNP from the construction of the (IBS-based) similarity matrix and rerunning the analysis with the remaining 25 SNPs. All of the analyses produced test statistics that were significant at the .005 level, which suggests that no SNP was solely responsible for the

Table 4. Distance-Based Regression-Analysis Results

			SNPs in HWE			All SNPs		
Distance Measure	Weighted	Haplotype	Pseudo-*F*	Permuted *P*[a]	Variation (%)	Pseudo-*F*	Permuted *P*[a]	Variation (%)
CHI3L2 allele sharing[b]	No	No	14.35	.00008	20.69	14.85	.00008	21.26
CHI3L2 rare-allele sharing[b]	Allele	No	12.30	.00021	18.28	12.08	.00019	18.01
CHI3L2 haplotype sharing[b]	No	Yes	16.40	.00007	22.97	14.52	.00008	20.88
CHI3L2 weighted allele sharing[b]	Functional	No	14.69	.00006	21.08	14.77	.00007	21.17
CHI3L2 weighted haplotype sharing[b]	Functional	Yes	14.80	.00005	21.2	13.46	.00008	19.66
CHI3L2 weighted allele sharing[b]	Conservation	No	14.35	.00008	20.69	14.85	.00008	21.26
CHI3L2 weighted haplotype sharing[b]	Conservation	Yes	9.78	.00008	15.09	8.87	.00008	13.89
CHI3L2 weighted allele sharing[b,d]	Association strength	No	1.12	.00001	15.54			
CHI3L2 weighted allele sharing[b]	Association strength	No	8.63	.00010	13.57			
CHI3L2 allele sharing,[b] only tSNPs	No	No	14.91	.00004	21.33			
CHI3L2 rare-allele sharing,[b] only tSNPs	Allele	No	15.98	.00003	22.51			
SQSTM1 allele sharing[c]	No	No	.8761	.3818	1.57			
SQSTM1 rare-allele sharing[c]	Allele	No	1.0224	.3848	1.83			
GSTM2 allele sharing[b]	No	No	.6768	.4824	1.22			
GSTM2 rare-allele sharing[b]	No	No	.8794	.4398	1.57			

[a] *P* value from 100,000 permutations in the similarity-matrix regression analysis.

[b] Association with averaged, $\log_2$ *CHI3L2* gene expression levels.

[c] Association with averaged, $\log_2$ *SQSTM1* gene expression levels.

[d] Only significant SNPs were used.

association signal, either because of LD relationships between the SNPs or because of independent functional effects of the SNPs. To minimize the LD relationships, we eliminated SNPs ($n = 14$) in complete LD ($D' = 1.0$) and found a significant association (table 5).

Second, we considered the effects of including nonassociated SNPs in the analysis by assigning individuals' alleles via a random-number generator. We constructed IBS-similarity matrices with the original *CHI3L2* SNPs plus these nonassociated SNPs. Figure 4 suggests that the original signal provided by the 26 *CHI3L2* SNPs was so strong that additional SNPs, comprising almost 80% of the SNPs used to construct the matrix, could not completely eliminate the statistical significance of the association. We note that the association signal steadily decreased the more we added nonassociated SNPs, which suggests that association strength can be used to identify sets of adjacent SNPs in a genomic region that influence phenotypic expression (fig. 4). Had we confined attention to only the nonassociated SNPs, no association would have been found. Essentially, the 26 SNPs in the *CHI3L2* gene were used initially to construct the similarity matrix. Additional SNPs that were randomly assigned to individuals—and hence were known not to be associated with the phenotype a priori—were added in greater numbers to those used to construct the similarity matrix. These matrices were then analyzed for association.

Third, we considered the analysis of simulated data generated in a few highly contrived settings involving interacting loci. We considered two biallelic polymorphic sites in a gene with alleles 0 and 1, for which the combination of alleles 0 and 0 at the two loci on a chromosome or the combination 1 and 1 (denoted as "0-0" and "1-1," respectively) raised the value of a phenotype by a value α. Combinations 0-1 and 1-0 decreased phenotype levels by α. We assumed that the 0 and 1 alleles at each locus were equally frequent. Note that the mean phenotype value of an allele at each locus would be 0.0. We then assumed five pairs of such interacting loci on a chromosome, for a total of 10 loci. We then randomly assigned 300 simulated individuals' alleles at the 10 loci on two chromosomes and derived their phenotype value by summing the appropriate 2-locus values over the 5 locus pairs. We added noise to their phenotype by generating a standard normal deviate and adding it to an individual's phenotype value. We considered α values of 0.5 (setting 1), 1.0 (setting 2), and 2.0 (setting 3). Note that single-locus analyses should not find associations with the phenotype in this setting, and, since we did not assume LD and there are 1,024 (i.e., 2^{10}) equally frequent possible haplotypes across the 10 loci, haplotype analyses that are based on phylogeny information or very extensive and somewhat arbitrary groupings are not likely to work. We subjected the simulated data (available on request) to standard single-locus analyses and the proposed similarity analysis based on IBS similarity. Standard single-locus analyses identified no significant results (table 6), whereas the similarity analysis produced *P* values of .029, .004, and .0024 for settings 1, 2, and 3, respectively.

Discussion

Genetic-association studies have been plagued with inconsistent results, so questions have been raised about phe-

Table 5. Effect on the Association Strength of Omitting Each *CHI3L2* SNP

The table is available in its entirety in the online edition of *The American Journal of Human Genetics.*

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure 4. The effect of including known nonassociated SNPs in the construction of the similarity matrix (IBS-similarity measure used).

nomena that may contribute to these inconsistent results.[1] Of the many factors that could create problems for large-scale genomewide association studies, those that concern data analysis are receiving a great deal of attention. Analytic methods are particularly difficult to assess and compare, since there are really no standards for judging them, given the many settings in which sequence variations can influence a phenotype. Thus, some methods may be better suited to one or a few of these settings than are others. The most basic approach to the analysis of sequence variation in association studies is to test each individual locus—independent of the other loci—for association with the trait or disease in question. This assumes that the effects of each locus, both within and across the genes studied, on phenotypic expression are independent. Although there is some research that considers the analysis of interactions both between and/or across different genes for association studies,[54] there is little research that considers the simultaneous effect of multiple variations *within* a gene. Thus, an alternative or complementary analysis approach for genetic associations would involve consideration of the actual *composition* of genes (i.e., consideration of the effects of particular combinations of variations in a gene that an individual possesses) and the impact that these multiple variations have on phenotypic expression. With this in mind, it is arguable that taking the more holistic view of genetic variation from a diplotype perspective may be appropriate. In addition, it is also arguable that future association studies should take advantage of strategies that exploit available biological knowledge about the functions of genes. Our proposed analysis strategy encompasses single-locus– and haplotype-phylogeny–based approaches to genetic association analyses, but it is much more flexible and has a number of advantages.

Advantages and Extensions of the Proposed Analysis Approach

The proposed association-analysis approach has many features that make it a good complement to existing analysis methods: it can accommodate many of the biological phenomena known to arise in human gene-phenotype relationships (e.g., humans are diploid, sequence variations do not work in isolation, etc.), it exploits the growing amount of information on sequence variations and their functional effects, it can be complemented with graphic aids to assist in interpretation of the data, and, unlike other association-analysis methods based on the assessment of similarity, it does not require cluster analysis.

In addition, one of the best features of the proposed method is its flexibility. The formulation of the test statistic can admit a wide range of analysis scenarios beyond analyses that focus on a single gene. For example, we are exploring the use of our procedure in the assessment of multiple genomic regions, using pathway information (authors' unpublished data; M. Zapala and N. J. Schork, unpublished data), the analysis of genome-scan data (J. Wessel, N. Malo, O. Libiger, and N. J. Schork, unpublished data), the analysis of multiple phenotypes (authors' unpublished data; N. J. Schork, J. Wessel, R. Salem, and D. T. O'Connor, unpublished data), and the analysis of genetic background (C. Nievergelt and N. J. Schork, unpublished data; M. Zapala and N. J. Schork, unpublished data). The procedure can also be used for the analysis of other data-analysis settings (e.g., the analysis of ecological data and gene expression data[44,50]) (M. Zapala and N.J.S., unpublished data).

Limitations of the Analysis Approach

There are a few limitations inherent in the proposed multilocus association–analysis approach. For example, predicting in vivo functional effects from in vitro studies can be problematic and, as such, may not provide appropriate weights for use in the construction of the similarity matrix. The same could be said of the use of model systems for providing insight into the physiologic effect of genomic variations in humans. In addition, the number of loci to include in the similarity calculations will not necessarily be known a priori, which is important if relevant SNPs are left out of the analysis or if too many irrelevant SNPs are used, which is a general problem with association-analysis methods and is not unique to the proposed approach. The interpretation of the conserved sequence surrounding SNPs can also be problematic. Many computational approaches to assessing conservation are limited simply by the available sequence information and the ability to align sequences from different species. Although the procedure critically depends on the choice of a similarity measure, this aspect of the procedure makes it appealing, since modeling the effects of genetic variations and comparing genomes can be pursued in a variety of ways, some of which may be more powerful in certain settings than in others. The power of the proposed approach in different analysis settings and locus-effect scenarios deserves attention, but, since our procedure is rooted in traditional ANOVA and regression modeling, many of the same intuitions and findings related to the power of these modeling procedures apply. For example, the proposed procedure assesses the question of how much of the variation in the similarity/dissimilarity exhibited

Table 6. Standard Regression Analysis–Based Single-Locus Results Involving the Data Generated with Interacting Loci

The table is available in its entirety in the online edition of *The American Journal of Human Genetics.*

by a group of individuals can be explained by another factor, which is analogous to questions concerning how much of the variation in a particular trait is explained by a certain factor in regression and ANOVA power-assessment contexts (authors' unpublished data).

Another issue with the proposed approach, which is an issue with all association-analysis methodologies, involves missing genotype data. One can handle missing genotype data in a number of ways. First, one could restrict the construction of the similarity measure to only those individuals with complete data—which may result in a substantially reduced sample size—or could simply construct the measure with the data that are available for each pair of subjects. This latter approach will be problematic if a number of individuals are missing genotype data at the most heavily weighted (i.e., functional) loci. Another approach to handling missing data would involve imputing or assigning individuals' genotype data on the basis of LD information. This approach would be only as useful as the strength of the LD between alleles at the loci with missing data and those with no missing data. The approach we took to handling missing data was to use whatever genotype information was available on the subjects for the similarity calculations. Since we had very little missing information (~1% of all genotype data was missing in the data set we used), we felt this approach was warranted.

A final issue of concern for association studies involves the effect of stratification or genetic-background heterogeneity. Our proposed association-analysis approach, like others, can accommodate such phenomena by simply including relevant covariates in the analysis (e.g., race, genetic background–cluster membership, degree of admixture, etc.) that reflect genetic-background information for the subjects in the study.

Despite limitations of the approach—which have less to do with the mechanics behind the approach and more to do with deficiencies in the available knowledge it tries to exploit—it is intuitive and flexible and can provide a complementary approach to existing methods for assessing multilocus data. The proposed approach is likely to have greater applicability and utility in a time when efficient and cost-effective sequencing technologies can be used to assess many individuals' genomes, since one can examine the similarity of these individuals' actual DNA sequences rather than examining commonality of sequence variations at a few well-chosen sites.

Acknowledgments

N.J.S. and his laboratory are supported in part by National Heart, Lung, and Blood Institute Family Blood Pressure Program research grant U01 HL064777-06; National Institute of Aging Longevity Consortium research grant U19 AG023122-01; National Institute of Mental Health Consortium on the Genetics of Schizophrenia research grant 5 R01 HLMH065571-02; National Institutes of Health research grants R01 HL074730-02 and R01 HL070137-01; and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator).

Web Resources

The accession number and URLs for data presented herein are as follows:

ESEfinder, http://rulai.cshl.edu/tools/ESE/
Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/ (for baseline expression levels of genes in CEPH individuals from the International HapMap Project [accession number GSE2552])
Gene Regulation, http://www.gene-regulation.com/pub/programs.html#pmatch (for P-Match)
GeneTree, http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html
HapMap, http://www.hapmap.org/
ITB Blast, http://www.ba.itb.cnr.it/BIG/Blast/BlastUTR.html (for BigBlast)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *CHI3L2*)
PipMaker and MultiPipMaker, http://pipmaker.bx.psu.edu/pipmaker/
PolyPhen, http://www.bork.embl-heidelberg.de/PolyPhen/
PupasView, http://pupasview.bioinfo.ochoa.fib.es/
RESCUE-ESE, http://genes.mit.edu/burgelab/rescue-ese/
SIFT, http://blocks.fhcrc.org/sift/SIFT.html
UTRScan, http://www.ba.itb.cnr.it/BIG/UTRScan/
VISTA Genome Browser, http://pipeline.lbl.gov/cgi-bin/gateway2
Vista Tools, http://genome.lbl.gov/vista/index.shtml (for rVISTA)

References

1. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61
2. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320
3. Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1: e32
4. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208
5. Hasler G, Drevets WC, Gould TD, Gottesman II, Manji HK (2006) Toward constructing an endophenotype strategy for bipolar disorders. Biol Psychiatry 60:93–105
6. Luo X, Kranzler HR, Zuo L, Wang S, Schork NJ, Gelernter J (2006) Diplotype trend regression analysis of the *ADH* gene cluster and the *ALDH2* gene: multiple significant associations with alcohol dependence. Am J Hum Genet 78:973–987
7. Small KM, Mialet-Perez J, Seman CA, Theiss CT, Brown KM, Liggett SB (2004) Polymorphisms of cardiac presynaptic α2C adrenergic receptors: diverse intragenic variability with haplotype-specific functional effects. Proc Natl Acad Sci USA 101: 13020–13025
8. Hamon SC, Stengard JH, Clark AG, Salomaa V, Boerwinkle E, Sing CF (2004) Evidence for non-additive influence of single nucleotide polymorphisms within the apolipoprotein E gene. Ann Hum Genet 68:521–535
9. Owen MJ (2005) Genomic approaches to schizophrenia. Clin Ther Suppl A 27:S2–S7
10. Weinshenker BG, Sommer S (2001) VAPSE-based analysis: a

two-phased candidate gene approach for elucidating genetic predisposition to complex disorders. Mutat Res 458:7–17

11. Levinson DF (2006) The genetics of depression: a review. Biol Psychiatry 60:84–92
12. Wang Z, Moult J (2003) Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vβ domain. Proteins 53:748–757
13. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17:263–270
14. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328
15. Krishnan VG, Westhead DR (2003) A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics 19:2199–2209
16. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612
17. Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region β 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97:10483–10488
18. Lee JH, Choi JH, Namkung W, Hanrahan JW, Chang J, Song SY, Park SW, Kim DS, Yoon JH, Suh Y, Jang IJ, Nam JH, Kim SJ, Cho MO, Lee JE, Kim KH, Lee MG (2003) A haplotype-based molecular analysis of CFTR mutations associated with respiratory and pancreatic diseases. Hum Mol Genet 12:2321–2332
19. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491
20. Gnanadesikan R, Kettering JR, Tsao SL (1995) Weighting and selection of variables for cluster analysis. J Classification 12: 113–136
21. Yu K, Gu CC, Province M, Xiong CJ, Rao DC (2004) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 27:182–191
22. Müller T, Selinski S, Ichstadt K (2005) Cluster analysis: a comparison of different similarity measures for SNP data (available at http://opus.zbw-kiel.de/volltexte/2005/3389/pdf/tr14-05.pdf) (accessed August 29, 2006)
23. Sielinski S (2005) Similarity measures for clustering SNP and epidemiological data (available at http://www.sfb475.uni-dortmund.de/berichte/tr25-06.pdf.pdf) (accessed September 19, 2006)
24. Trooskens G, De Beule D, Decouttere F, Van Criekinge W (2005) Phylogenetic trees: visualizing, customizing and detecting incongruence. Bioinformatics 21:3801–3802
25. Kibbey C, Calvet A (2005) Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. J Chem Inf Model 45:523–532
26. Hughes T, Hyun Y, Liberles DA (2004) Visualising very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinformatics 5:48
27. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1:274–286
28. Mountain JL, Cavalli-Sforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. Am J Hum Genet 61:705–718
29. Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766
30. Belkhir K, Castric V, Bonhomme F (2002) IDENTIX, a software to test for relatedness in a population using permutation methods. Mol Ecol Notes 2:611–614
31. Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel EW, Rubin EM, Hamann B, Dubchak I (2004) Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. Bioinformatics 20:636–643
32. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–W279
33. Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglou S, Dubchak I (2004) Automated whole-genome multiple alignment of rat, mouse, and human. Genome Res 14:685–692
34. Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351
35. Seltman H, Roeder K, Devlin B (2003) Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25: 48–58
36. Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. Genetics 156:1259–1275
37. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390
38. Seltman H, Roeder K, Devlin B (2001) Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:1250–1263
39. Halperin E, Eskin E (2004) Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics 20: 1842–1849
40. Excoffier L, Laval G, Schneider S (2005) Arlequin ver 3: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50 (http://www.la-press.com/evolbio05.htm) (accessed August 31, 2006)
41. Page RD (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819–820
42. Humphreys K, Iles MM (2005) Fine-scale mapping in case-control samples using locus scoring and haplotype-sharing methods. BMC Genet Suppl 6:S74
43. Bardel C, Darlu P, Genin E (2006) Clustering of haplotypes based on phylogeny: how good a strategy for association testing? Eur J Hum Genet 14:202–206
44. Edgington ES (1995) Randomization tests. Marcel Dekker, New York
45. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–338
46. McArdle BH, Anderson MJ (2001) Fitting multivariate models

to community data: a comment on distance-based redundancy analysis. Ecology 82:290–297

47. Good PI (2000) Permutation tests. Springer Verlag, New York
48. Manly B (1997) Randomization, bootstrap, and Monte Carlo methods in biology. Chapman and Hall, London
49. Jockel KH (1986) Finite sample properties and asymptotic efficiency of Monte Carlo tests. Ann Stat 14:336–347
50. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46
51. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369
52. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769
53. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1–12
54. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417

Online Tables.

Table 5. Effect on the Association Strength of Omitting Each CHI3L2 SNP.

SNP(s) Removed	Pseudo-F	Permutation P	Variation (%)
ALL 26 SNPs	14.35	0.00008	21
rs755467	13.36	0.00019	20
rs2147790	15.93	0.00007	22
rs2255089	15.40	0.00008	22
rs2274232	15.56	0.00007	22
rs2147789	15.13	0.00008	22
rs2182115	15.52	0.00007	22
rs1325284	14.50	0.00008	21
rs2251715	15.40	0.00008	22
rs961364	13.57	0.00017	20
rs2764543	14.62	0.00007	21
rs7366568	14.99	0.00007	21
rs2820087	14.58	0.00008	21
rs6685226	15.54	0.00007	22
rs11583210	16.44	0.00007	23
rs12032329	15.42	0.00008	22
rs2477578	14.50	0.00008	21
rs2494006	14.55	0.00008	21
rs7542034	15.30	0.00007	22
rs942694	14.61	0.00008	21
rs942693	14.50	0.00008	21
rs2182114	14.50	0.00008	21
rs5003369	14.50	0.00008	21
rs11102221	16.50	0.00007	23
rs3934922	13.46	0.00018	20
rs3934923	14.50	0.00008	21
rs8535	13.36	0.00019	20
Correlated SNPs[a]	9.82	0.00026	15

[a] Correlated SNPs were removed, and preference was given to functional SNPs (correlation or D’=1; deleted SNPs 2, 3, 4, 9, 10, 11, 16, 17, 19, 20, 21, 22, 25, and 26).

Table 6. Standard Regression Analysis-Based Single-Locus Results Involving the Data Generated with Interacting Loci

	P for Setting					
	1		**2**		**3**	
SNP	**exact**	**corrected**	**exact**	**corrected**	**exact**	**corrected**
1	0.1806	0.8636	0.1443	0.7895	0.1511	0.8057
2	0.6351	1.0000	0.6355	1.0000	0.6648	1.0000
3	0.4791	0.9985	0.6972	1.0000	0.9260	1.0000
4	0.2211	0.9178	0.4190	0.9956	0.6914	1.0000
5	0.6838	1.0000	0.9807	1.0000	0.7683	1.0000
6	0.0298	0.2611	0.1273	0.7438	0.3883	0.9927
7	0.3276	0.9811	0.3488	0.9863	0.4107	0.9949
8	0.9239	1.0000	0.8635	1.0000	0.8254	1.0000
9	0.3952	0.9935	0.3996	0.9939	0.4447	0.9972
10	0.1948	0.8854	0.8968	1.0000	0.3980	0.9937

NOTE.--Raw data used in the analyses are available from the authors. Corrections are based on the Sidak correction for multiple comparisons.

Online Figures.

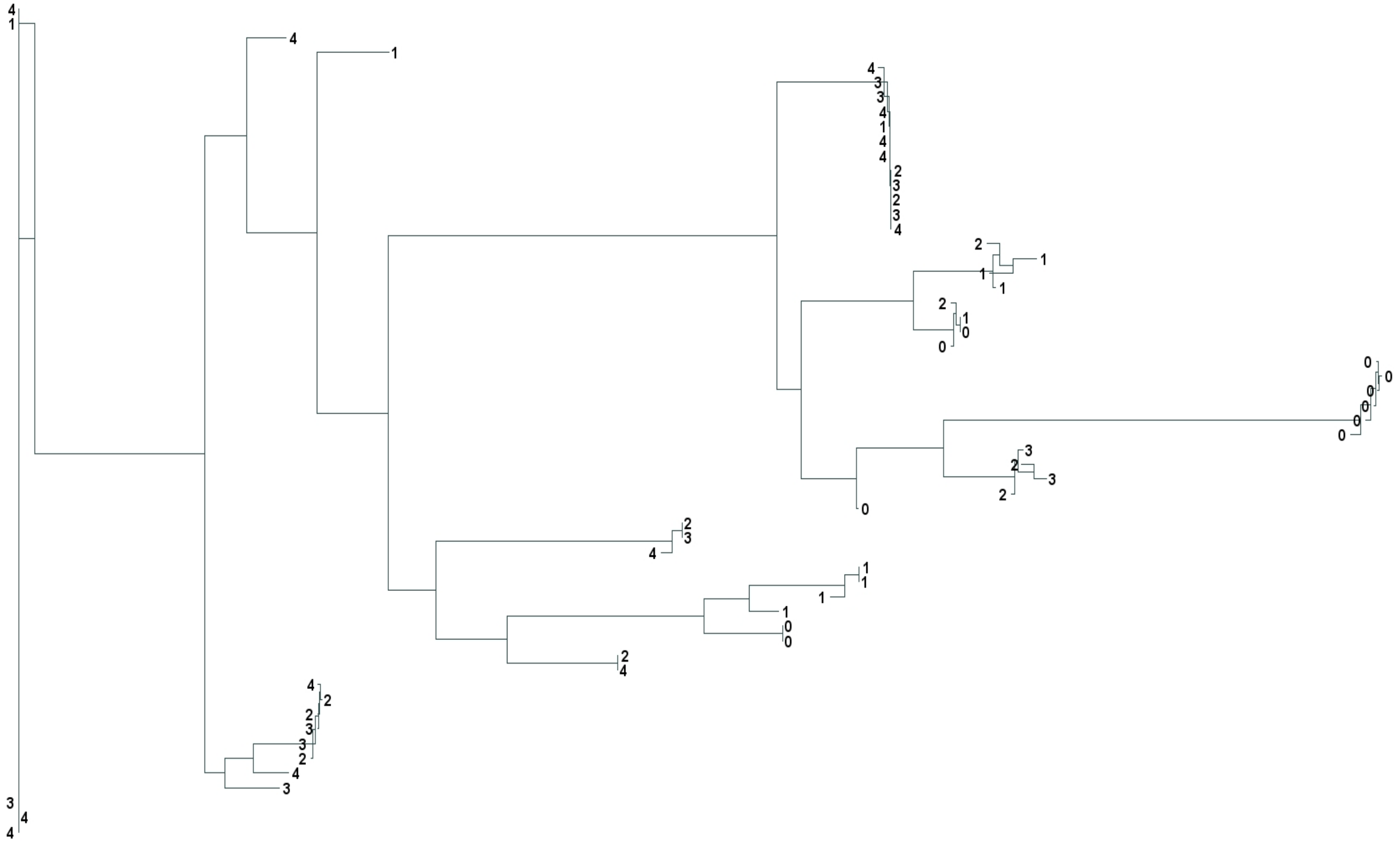


Figure 2. Tree representation of the similarity in the allelic profiles of 57 unrelated CEPH individuals based on variations in the CHI3L2 gene with use of a standard IBS allele-sharing measure. The proximity of the branches associated with individuals corresponds to greater similarity. Individual branches are coded such that, the larger the number, the greater the CHI3L2-expression value an individual has. It can be seen that individuals with similar number codes (i.e., individuals with similar CHI3L2-expression values) are on branches near each other (in general), which is indicative of the association between greater CHI3L2 allelic–composition similarity and CHI3L2 gene–expression level.

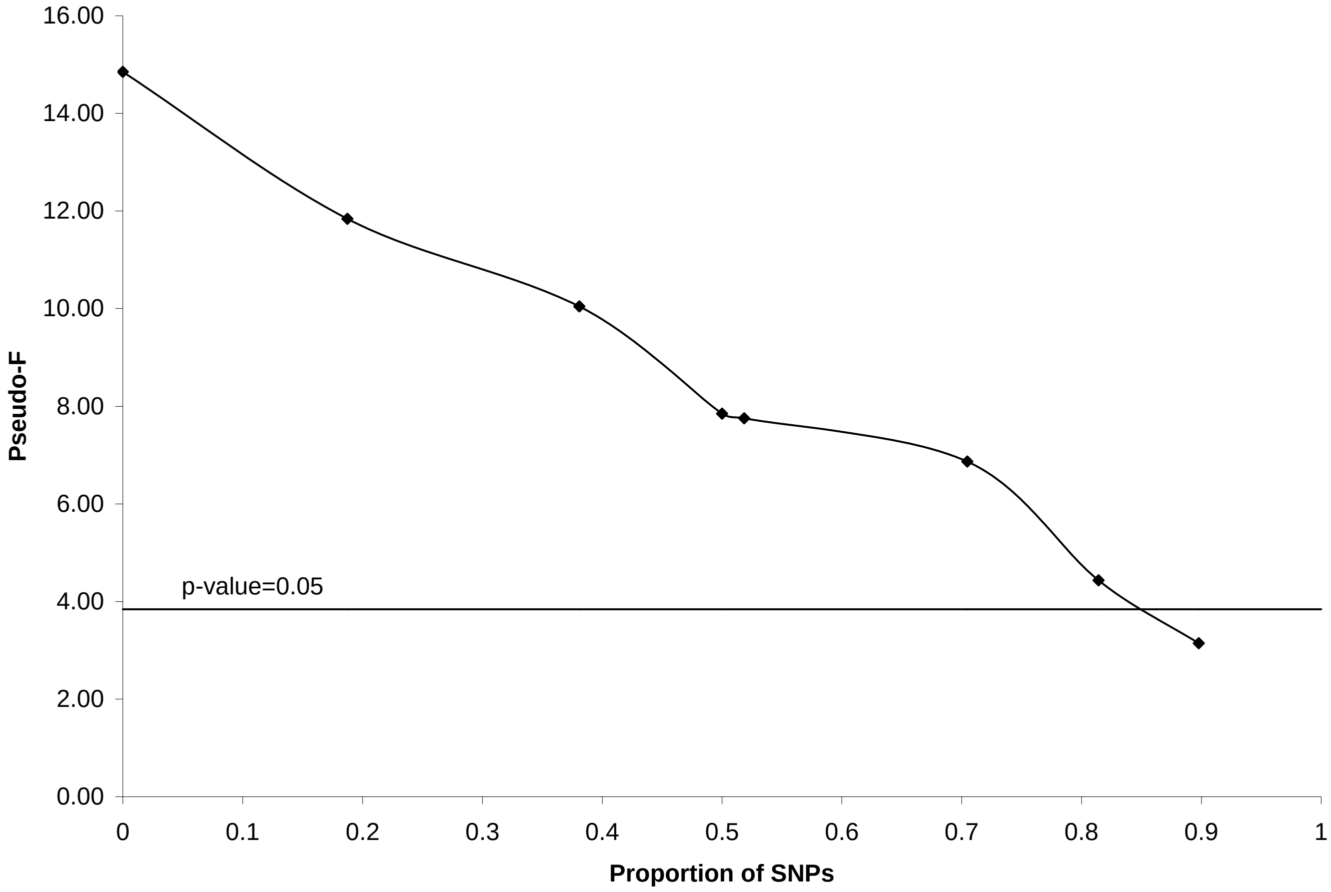


Figure 4. The effect of including known nonassociated SNPs in the construction of the similarity matrix (IBS similarity measure used).

ACKNOWLEDGEMENTS

The text of Chapter 2 has been published in the American Journal of Human Genetics (79(5):792-806, November, 2006) for publication as:

> Jennifer Wessel, and Nicholas J Schork. Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis

The dissertation author was the primary researcher and/or author and the co-author listed in this publication directed and supervised the research which forms the basis for this chapter.

CHAPTER 3

Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis

ABSTRACT

Emerging high-throughput genotyping technologies and the extensive characterization of variation in the human genome has made genome wide association studies a reality. However, a number of very complex issues surround the analysis of, e.g., >500,000 Single Nucleotide Polymorphisms (SNPs) in such studies. We describe the utility of an analysis approach, termed Multivariate Distance Matrix Regression (MDMR), in such studies. The MDMR has previously been shown to possess some very desirable properties for candidate gene studies, such as its very comprehensive and flexible nature. We apply the MDMR approach to publicly available CHI3L2 gene expression data as well as 811,886 phased SNPs from the International HapMap Project to showcase the utility of the method. We implement the MDMR analysis approach in a “moving windows” analysis strategy, where the window sizes included 1, 2, 5, 10, 20 and 50 SNPs across all 22 autosomes. We identified a number of potentially interesting loci in our analysis and describe the genes within the associated regions. These genes are known to be involved in a wide variety of biological functions, most notably DNA binding and transcription. We also describe limitations of the proposed approach as well as areas for future research.

INTRODUCTION

With the completion of the International HapMap Project [1], and increasing efficiencies in genotyping technologies that can accommodate a large number of polymorphic loci (e.g., Single Nucleotide Polymorphisms or SNPs) in single chip-based assays, genome-wide association (GWA) studies are now a reality for researchers [2-4]. GWA studies essentially involve testing a large enough number of genotyped polymorphic sites for association with a trait or disease information on an appropriate sample of individuals. The ultimate goal is to genotype enough markers to cover the genome in its entirety. The 300,000–1,000,000 SNPs that can be interrogated on current SNP genotyping chips that will be used in GWA studies, however, pose many difficult analysis problems for researchers, not the least of which is controlling for the very large number of multiple comparisons and statistical tests performed in such studies. One of the most widely used approaches to GWA analysis involves testing each SNP individually for association with a phenotype. This approach is problematic for many reasons, as it ignores the perhaps weak linkage disequilibrium (LD) that may exist between any genotyped locus and a trait-relevant locus at an adjacent locus and the fact that many loci in a given genomic region may work in tandem to influence the expression of the trait.

There are other issues that will inevitably arise in WGA studies that have been known to plague simple candidate gene association studies as well, such as population stratification, phenotypic heterogeneity, and biological meaningfulness of the association results [5-11]. Other issues that create unique problems in WGA studies, for example, involve the coverage of the genome associated with any standardized SNP genotyping sets of the type represented on available genotyping chips [12].

One of most vexing and difficult problems with WGA studies involves the models and statistical methods used to relate genotype and/or haplotype information to the phenotype of interest. Although a number of analysis methods are being developed for WGA studies, such as Bayesian graphical models and likelihood weighting schemes [13, 14], and methods to improve power using prior linkage information [15], that may perform better than standard single locus-oriented analyses, there is as of yet no real consensus on the best way to approach WGA study analysis.

We previously developed a regression method that involves the consideration of multiple loci in an association analysis through the derivation of measures of genomic similarity. These measures of genomic similarity can incorporate weighting schemes based on prior knowledge about the biological effects of the SNPs tested, prior association results, haplotype phylogeny information, and a host of other factors. In addition, the method can be used to explicitly account for the diploid nature of the human genome. We considered appropriate test statistics for relating the measures of genomic similarity to a phenotype as well as the repeated use of these statistics in genetic mapping studies [16]. This method can be termed the ‘Multivariate Distance Matrix Regression’ or ‘MDMR’ analysis approach for genetic association studies.

To showcase the MDMR method in WGA settings, we pursued an analysis of over 800,000 SNPs collected on CEPH individuals in the HapMap database (www.hapmap.org) who had been assayed for CHI3L2 gene expression phenotypes (as well as many other gene expression phenotypes) obtained from immortalized lymphocytes collected from them. We chose to consider CHI3L2 gene expression because this gene's expression levels were found to be influenced significantly by a single SNP within an intronic region of the gene itself (Cheung et al. 2005), and we have considered a more detailed analysis of this gene in a prior publication [16]. We also note that molecular phenotypes, such as gene expression levels, are becoming important diagnostic tools for treatment and disease outcome, which raises questions about their relationships to naturally occurring DNA sequence variation [17]. To implement the MDMR procedure in WGA analysis settings we consider a moving window based strategy in which some number, *k*, of adjacent SNPs are used in the computation of a similarity matrix, the matrix is then tested for an association between variations at the *k* loci and the phenotype, and then the window is moved one SNP away and the analysis is repeated. The choice of the window size is arbitrary, but can be varied in order to identify SNP effects that appear to work in aggregate or in isolation, thus allowing for flexibility in the analysis.

METHODS

Phenotype and Genotype Data

We obtained SNP data collected on 57 unrelated CEPH individuals from the International HapMap Project database (www.hapmap.org). These individuals were chosen by HapMap researchers for massive, genome-wide genotyping studies [1] and also used to assess gene expression patterns obtained from immortalized lymphocytes [2]. We downloaded the gene expression data as it is publicly available (via GEO accession number: GSE2552; http://www.ncbi.nlm.nih.gov/geo/). Our analyses excluded the individual labeled NA06993 in the gene expression studies because detailed analysis of HapMap data suggested that the sample associated with this person is likely to have derived from an unreported relative. We also added data associated with individual labeled NA12056 since gene expression data for this individual is now available. We ultimately downloaded phased, haplotype data on the 22 autosomal chromosomes from the HapMap (phase 1) database that were genotyped on the 57 CEPH individuals. We focused on the CHI3L2 gene expression phenotype, since, as mentioned previously, Cheung et al. (2005) found linkage and cis-acting variations influencing this gene's expression and also demonstrated functional evidence for an intronic SNP in CHI3L2 [2]. Monomorphic SNPs were eliminated from analyses leaving 811,886 SNPs for analysis. Repeated, multiple gene expression values collected on each of the CEPH individuals were averaged when available. We considered use of $\log_2$-transformed expression levels due to skewness in the expression values.

MDMR Analysis

We applied the MDMR analysis method described by Wessel and Schork [16] to SNPs across the genome and the CHI3L2 gene expression phenotype. In brief, the MDMR method involves the characterization and measurement of the similarity/dissimilarity of the allelic composition of a set of individuals' diploid genomes in a region of interest. For the present study we used a simple identity-by-state (IBS) measure of haplotype pair (or 'diplotype') similarity [16]. The resulting matrix was then tested for patterns consistent with their being an association between the SNPs in the region of interest and the phenotype using an F-statistic [16]. We defined a 'region' using a moving window strategy. In our initial analyses, we used a window size of 10, scrolling across each chromosome using 10 adjacent SNPs to calculate haplotpye similarity. We repeated the analyses using window sizes of 1, 2, 5, 10, 20 and 50 on chromosome 1 (i.e., the chromosome the CHI3L2 gene is located on). We identified SNPs in the analysis within 100kb around the CHI3L2 gene so that we could determine if our analyses could effectively 'recover' the known association between variations in the CHI3L2 gene and CHI3L2 gene expression.

We also pursued analyses that considered each locus in isolation using the proposed similarity regression procedure in a more detailed analysis of SNPs within 3 megabases (MB) around the CHI3L2 gene (base positions: 109,568,013 – 112,588,629 based on the latest release of the human genome via the Golden Path website (http://genome.ucsc.edu/). To more accurately assess the significance of associations (instead of merely relying on the putative asymptotic distribution of the proposed F-statistic under the null hypothesis of no association) we used 100,000 permutations of the data to assess the probability of a type I error, with the exception of our more detailed analysis of SNPs around the CHI3L2 gene for which we used 1000 permutations.

Graphical Display of Similarity Matrices

Similarity matrices can be represented graphically in a number of ways that can facilitate interpretation. We consider 'heatmaps' which simply color code the elements of a similarity matrix such that higher similarity values are represented as 'hotter' or more red colors and lower similarity values are represented as 'colder' or more blue colors [18]. If the matrix is ordered such that individuals with similar phenotype values are next to each other, then neighboring cells along the diagonal of the matrix (representing individuals with similar phenotype values) will present patches of red, indicating a relationship between the phenotype values and genomic similarity (see Figures 1a and 1b, discussed in the results section).

RESULTS

Basic WGA analysis

We analyzed SNPs on the 22 phased autosomes using 10 locus windows and the MDMR analysis technique. Since the CHI3L2 gene is located on chromosome 1, we used the peak F-statistic associated in this region as our F-statistic threshold for exploring other chromosomal regions likely to possess

CHI3L2 gene expression-influencing loci (Figure 2, peak pseudo-F statistic = 16.11; p=0.00001). Using this threshold, we identified additional peaks on chromosomes 2, 7, 10 and 12. We found, however, a number of peaks on each of these chromosomes (Table 1 and Figure 3). We explored the genes known to reside in the regions of the chromosomes and found that the genes under these peaks are involved in a number of biological functions that make sense with respect to the phenotype we studied, e.g. transcription regulators, while some genes had unknown functions or were only hypothetical proteins whose functions have yet to be determined.

In addition to the peak on chromosome 1 in the CHI3L2 gene region, chromosome 10 yielded the highest peak (figure 3, pseudo-F=18.80, p=0.00004). The SNPs located within the regions of this peak were located in or near the WAC gene, which contains a WW domain known to be involved in signal transduction. In a region of chromosome 2, SNPs showing association were in the FLJ16124 gene (figure 3, pseudo-F statistic = 16.69; p=0.00006), and in a chromosome 12 region, SNPs near the peak associated marker were in the 3' UTR of the LGR5 gene. For chromosome 7, the strongest associated SNPs involved a locus >480kb downstream from the ANKRD7 gene, but two other peaks on chromosome 7 involved SNPs that are known transcription regulators, NEUROD6 and CUTL1. The majority of the associations were in genic regions and were located in promoters, UTRS or introns, except the one >480kb away from a gene, ANKRD7. This result – based on the biology of the variations coupled with the statistical analysis – makes the results more compelling, since our phenotype, the expression level of the CHI3L2 gene, is likely influenced by, e.g., transcription factors and associated genes. We do, however, note that the SNPs that appeared to be associated with CHI3L2 gene expression that were more than >480kb from a gene on chromosome 7 were not in LD with SNPs in the nearest gene, ANKRD7. Further investigation showed that the SNPs were between ANKRD7 and hypothetical protein, LOC646752 (NCBI Map Viewer) but still out of the genic region (>100kb).

Variable Window Size Analyses

Table 2 summarizes the SNPs with the largest association identified through the use of different window sizes across chromosome 1. We want to emphasize that the F-statistic used in our analysis has the same number of degrees of freedom no matter how many SNPs are used to construct the measure of genomic similarity. Thus, analyses with different window sizes are comparable. Many of the associations were found in the region of the CHI3L2 gene and most were the strongest associations identified in all of our analyses. Each window size picked up the CHI3L2 region, based on our previously defined threshold and permutation analyses, with the exception of the window size of 50, suggesting that this window size was likely containing too many variations that did not have potential biological effects on CHI3L2 expression and hence were saturating or reducing the signals provided by the SNPs with biological effects.

We did find some associations that appeared only to be significant only when a particular window size was used. These included SNPs in PEX14 and

ACP6 genes (window size of 1), and the LMX1A gene, a transcription factor (window size of 2). Many associations were picked up in multiple window sizes, however, such as associations involving SNPs in the DEGS1 gene, where the SNP from window size 1 is contained in the other window sizes that appeared significant, as well as the PAP2D gene. It is also of interest to note that associations in SNPs in the CHI3L2 gene itself appeared to be strongest in different regions in the gene based on the window size used; e.g. the 2 SNPs with window size of 1 that happen to be in separate LD blocks based on the HapMap data, as well as the two associations involving the RAP1A gene. Some of these associations, however, were not identified in our initial 10 locus window analysis because they did not meet our pseudo-F threshold of 16.11. We examined single-locus associations involving SNPs 3 MB around the CHI3L2 gene using the haplotype similarity measure (Figure 4). For the 953 SNPs that were in this region, 90 (9.4%) SNPs were significant at $p \leq 0.05$ (based on 1000 permutations) of which 17 (18.9%) were in the actual defined CHI3L2 region. Another peak association was in the RAP1A gene, involved in GTPase activity, an association observed using windows of size 1 and 2.

In general, we found that as the window size is increased, such as with window size of 50, the results become more unpredictable, since the SNPs within these windows may span different genes that do not appear to have an obvious function associated with CHI3l2 expression. In general, the larger the window size, the smaller the F-statistic values. However, in theory, the use of a large window size may help identify regulatory blocks that do indeed involve many genes or regulatory factors (e.g., enhancers, silencers, microRNA genes, etc.). Thus, a window-size of 1 may provide greater potential for false-positive results, while a large window size, such as 50, may provide greater potential for false-negative results, but this may not be the case in certain settings.

Assessing Association Signal Strength

Because of the observation that greater window sizes appeared to produce less strong association signals, most likely because non-associated SNPs in large windows deplete the signal associated with SNPs within those windows, we considered the effects of the inclusion of known non-associated SNPs in an analysis through the use of SNPs in regions throughout the genome that were not found to be associated with CHI3L2 gene expression. We simply included these SNPs along with SNPs in regions that did show association with CHI3L2 gene expression in MDMR analyses to assess the effect of their inclusion in an analysis. Thus, we constructed haplotype similarity matrices with the original CHI3L2 SNPs plus these non-associated SNPs. We found that the addition of as many as 90% non-associated SNPs was needed before the association was not significant (i.e., produced a $p>0.05$ in the present analyses). We note that the association signal steadily decreased when we added more and more non-associated SNPs, suggesting that association strength can be used to identify sets of adjacent SNPs in a genomic region that influence phenotypic expression (see Figure 5). Had we confined attention to only the non-associated SNPs, no association would have been found.

DISCUSSION

We have identified a number of potential genomic sites involved in the expression of the CHI3L2 gene using a moving window-based MDMR analysis strategy. Many of these sites have logical biological functions, such as transcription. The ability to compare and contrast results obtained with analyses involving different window sizes provides flexibility in the analysis. Based on our analyses, however, it would appear that it may be best to use a medium window size (such as 10 or 20) – although this may be dependent on sample size, something we did not investigate – to identify regions of interest initially, and then pursue analyses for the regions that emerge as most significant using smaller window sizes or single-locus analyses. Merely using single locus analyses in a WGA analysis and not considering the biological basis of the putative association is likely going to lead to false positive results [13].

There are some limitations to our proposed analysis strategy. For example, we confined attention to the use of phase 1 SNPs, which does not represent all of the SNPs that could have been studied now that some 10,000,000 polymorphic sites have been identified in the genome. Some of these SNPs – although not genotyped on the individuals in our study – may be of direct relevance to CHI3L2 gene expression. In addition, some regions of chromosomes were not well genotyped by the HapMap researchers for various reasons, so that key genomic regions may not have been interrogated in our analyses. Also, some associations were found to involve SNPs that reside outside of a gene region. Although these SNPs could be in strong LD with other

SNPs in important genes this requires further work. Using the NCBI MapViewer (http://www.ncbi.nlm.nih.gov/mapview/), we were able to identify genes that were not viewable or available on the HapMap website, suggesting that there is more information that one could work with if additional genotyping was to be pursued with the CEPH subjects to explore CHI3L2 gene expression.

In terms of the MDMR analysis procedure itself, we want to emphasize that other groups are beginning to employ multiple locus analyses in an effort to capture effects not likely detectable in single-locus association analyses [13]. It may be the case that the use of an allelic composition-based similarity metric in WGA analysis is simply not as powerful – either in total or with respect to certain settings – than other analysis approaches. This issue demands further attention and can be pursued by comparing methods to validating association analysis data or through the use of simulation study methods (Malo, Wessel, and Schork, manuscript in preparation). The measure of similarity used is also an issue with MDMR analyses and greater attention needs to be given to the choice of a metric. We chose to use the most basic method – simple IBS allele sharing across haplotypes – but others, for example those that use weighting schemes, may be the most appropriate and powerful [16]. Finally, although the MDMR analysis procedure can explicitly account for allelic heterogeneity, we did not consider analyses that consider locus heterogeneity. Clearly, methods that leverage the combined effects of multiple genomic regions are as important in the analysis of complex human traits and diseases as are those that consider the biological reality of multiple variations with a single region.

REFERENCES

1. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437: 1299-320

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

3. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78

4. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38: 617-9

5. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429: 446-52

6. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6: 109-18

7. Lawrence RW, Evans DM, Cardon LR (2005) Prospects and pitfalls in whole genome association studies. Philos Trans R Soc Lond B Biol Sci 360: 1589-95

8. Todd JA (2006) Statistical false positive or true disease pathway? Nat Genet 38: 731-3

9. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243-6

10. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, et al. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38: 556-60

11. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33: 518-21

12. Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38: 659-62

13. Verzilli CJ, Stallard N, Whittaker JC (2006) Bayesian graphical models for genomewide association studies. Am J Hum Genet 79: 100-12

14. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38: 663-7

15. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78: 243-52

16. Wessel J, Schork NJ (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79: 792-806

17. Wessel J, Zapala M, Schork NJ (2006) Accommodating Pathway Information in Expression Quantitative Trait Locus (“eQTL”) Analysis. Genomics Submitted.

18. Kibbey C, Calvet A (2005) Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. J Chem Inf Model 45: 523-32

Table 1. MDMR Whole Genome Association Results Based on a Moving Window Size of 10 Adjacent Loci.

Chr	pseudo-F	p-value	var %	position	rs#	in a gene?*	gene name	function	location
1	16.11	0.00001	22.7	111087506	rs3886706	Yes	CHI3L2	Extracellular matrix protein	3' UTR (3kb)
1	16.06	0.00015	22.6	98992749	rs6701980	Yes	PAP2D	Hydrolase activity	5' (59kb)
1	15.78	0.00016	22.3	111077825	rs2764543	Yes	CHI3L2	Extracellular matrix protein	intron 7
1	12.01	0.00030	17.9	209493940	rs351407	Yes	PPP2R5A	Protein serine/threonine phosphatase activity	5' UTR (21kb)
2	17.00	0.00006	23.6	65638627	rs1437465	Yes	FLJ16124	Integral membrane protein	intron
2	13.60	0.00015	19.8	183482837	rs10497600	Yes	PDE1A	Phosphoric diester hydrolase activity	intron
2	14.73	0.00022	21.1	26084886	rs3845683	Yes	ASXL2	DNA binding	3' UTR (9kb)
7	16.71	0.00020	23.3	117936527	rs2689740	No	NA	NA	3' from ANKRD7 (>500KB)
7	11.76	0.00029	17.6	31119641	rs10238918	Yes	NEUROD6	Transcription regulator activity	3' UTR
7	11.80	0.00103	17.7	101419231	rs407943	Yes	CUTL1	Transcription regulator activity	intron
10	18.80	0.00004	25.5	28916977	rs332184	Yes	WAC	Signal transduction	3'UTR (3kb)
10	13.88	0.00019	20.2	50367536	rs4838566	Yes	OGDHL	unknown	5' (53kb)
10	13.68	0.00086	19.9	15883447	rs7903095	Yes	C10orf97	Cell cycle control	intron
12	16.21	0.00016	22.8	70270822	rs10879305	Yes	LGR5	G-protein coupled receptor activity	3' UTR (6kb)
12	14.21	0.00022	20.5	122762601	rs7133378	Yes	CCDC92	Unknown	3' UTR (11kb)
12	14.45	0.00035	20.8	122823432	rs7311969	Yes	ZNF664	Transcription regulator activity	UTR

Key: Chr=chromosome, var=variation, NA=not applicable. Genomic position and rs# are for the fifth marker. *100KB around the gene.

Table 2. MDMR Association Analysis Results According to Window Size for Chromosome 1.

window size	pseudo-F	p-value	variation	position	rs#	in a gene?	gene name	location	function
1	30.71	0.00001	35.8	111069150	rs755467	Yes	CHI3L2	intron 1	Extracellular matrix protein
1	24.01	0.00005	30.4	111396393	rs2800901	Yes	RAP1A	intron	GTPase activity
1	22.93	0.00003	29.4	144655381	rs10494246	Yes	ACP6	3' UTR (6kb)	Acid phosphatase activity
1	22.36	0.00003	28.9	111092960	rs7537675	Yes	CHI3L2	3' UTR (8kb)	Extracellular matrix protein
1	20.05	0.00008	26.7	10382740	rs479407	Yes	PEX14	intron	Signal transduction
1	19.97	0.00008	26.6	111090371	rs942696	Yes	CHI3L2	3' UTR (6kb)	Extracellular matrix protein
1	19.59	0.00005	26.3	221350016	rs11589025	Yes	DEGS1	intron	Metabolisim
2	37.61	0.00001	40.6	111084194	rs3934923	Yes	CHI3L2	intron 10	Extracellular matrix protein
2	19.69	0.00008	26.4	98986970	rs10747502	Yes	PAP2D	5' (54kb)	Hydrolase activity
2	18.13	0.00003	24.8	111392716	rs7553961	Yes	RAP1A	intron	GTPase activity
2	17.95	0.00008	24.6	221348239	rs4653996	Yes	DEGS1	intron	Metabolisim
2	15.18	0.00017	21.6	162441382	rs883864	Yes	LMX1A	intron	Transcription factor activity
5	26.22	0.00001	32.3	111090230	rs942697	Yes	CHI3L2	3' UTR (5kb)	Extracellular matrix protein
5	18.26	0.00010	24.9	98983465	rs1350177	Yes	PAP2D	5' (50kb)	Hydrolase activity
5	13.76	0.00026	20.0	221338001	rs6426178	Yes	DEGS1	5' (8kb)	Metabolisim
5	13.50	0.00045	19.7	209493940	rs351407	Yes	PPP2R5A	5' (11kb)	Protein serine/threonine phosphatase activity

Table 2. Continued.

window size	pseudo-F	p-value	variation	position	rs#	in a gene?	gene name	location	function
10	16.11	0.00001	22.7	111084194	rs3934923	Yes	CHI3L2	intron 10	Extracellular matrix protein
10	16.06	0.00015	22.6	98983465	rs1350177	Yes	PAP2D	5' (50kb)	Hydrolase activity
10	15.78	0.00016	22.3	111069150	rs755467	Yes	CHI3L2	intron 1	Extracellular matrix protein
10	12.01	0.00030	17.9	209482269	rs351377	Yes	PPP2R5A	5' (33kb)	Protein serine/threonine phosphatase activity
20	15.54	0.00006	22.0	111068887	rs7554451	Yes	CHI3L2	promoter (171bp)	Extracellular matrix protein
20	11.70	0.00038	17.5	98962946	rs12091525	Yes	PAP2D	5' (30kb)	Hydrolase activity
20	10.82	0.00037	16.4	209435897	rs7546833	Yes	PPP2R5A	5' (79kb)	Protein serine/threonine phosphatase activity
50	4.00	0.03732	6.8	150944424	rs4073768	Yes	INTS3	intron	Unknown
50	3.63	0.02633	6.2	150026996	rs6674372	Yes	LCE1A	3' (10kb)	Structural molecule activity
50	1.45	0.23609	2.6	80081648	rs10493661	No	NA	>2MB from a gene	NA

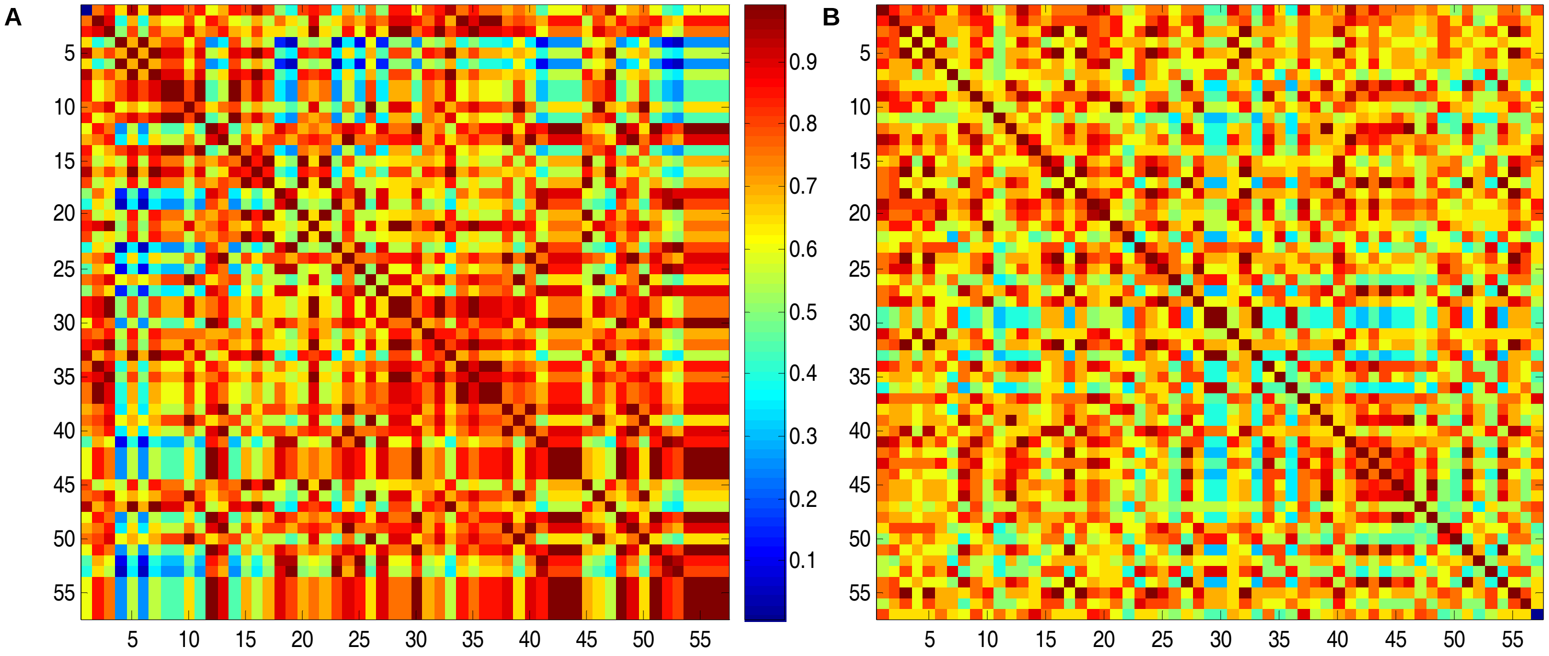


Figure 1. Heatmap representations of the similarity in the allelic profiles of 57 unrelated CEPH individuals based on variations in the WAC gene using the haplotype sharing measure (panel A; see text for details), and based on variations in the IL10 gene using the haplotype sharing measure (panel B). Note that individuals have been ordered in the matrix by increasing CHI3L2 levels. The concentration of "red" cells in the matrix along the diagonal in panel A suggests an association between similarity in the WAC gene composition and CHI3L2 expression. The lack of a pattern in panel B suggests that no association between similarity in IL10 gene composition and CHI3L2 expression exists.

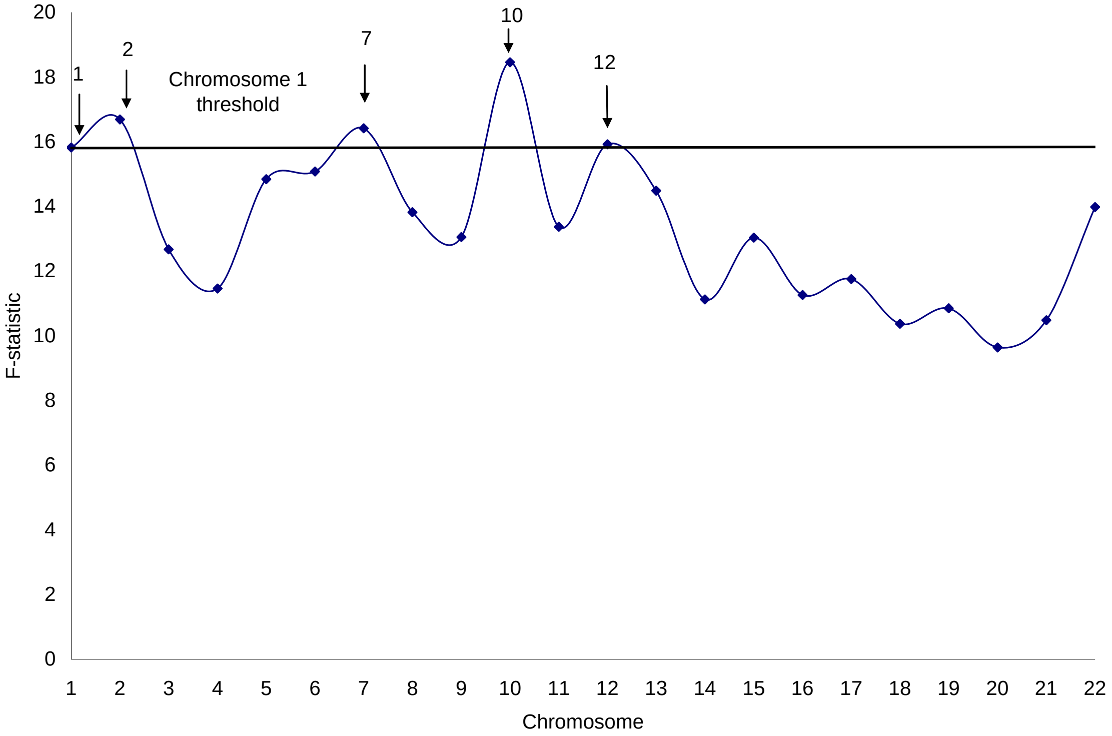


Figure 2. The peak associations from each of the 22 autosomes of 57 unrelated CEPH individuals. The peak association on chromosome 1, representing SNPs in the CHI3L2 gene, was used as the threshold for exploring other regions in the genome. These regions included peaks on chromosomes 2, 7, 10 and 12.

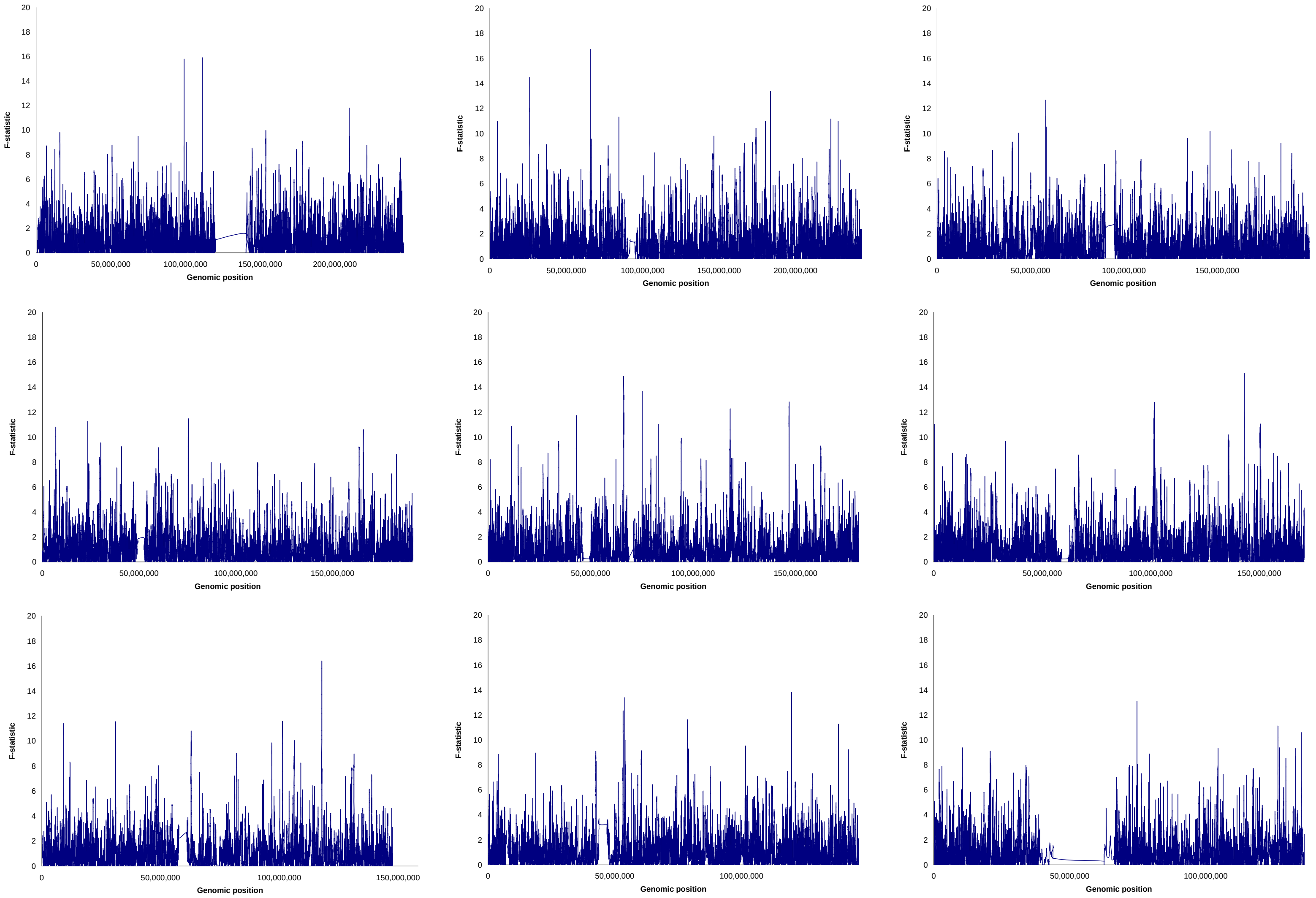


Figure 3. Associations for each of the 22 autosomes using 811,886 SNPs and a moving window size of 10. The chromosomes are in numerical order from left to right.

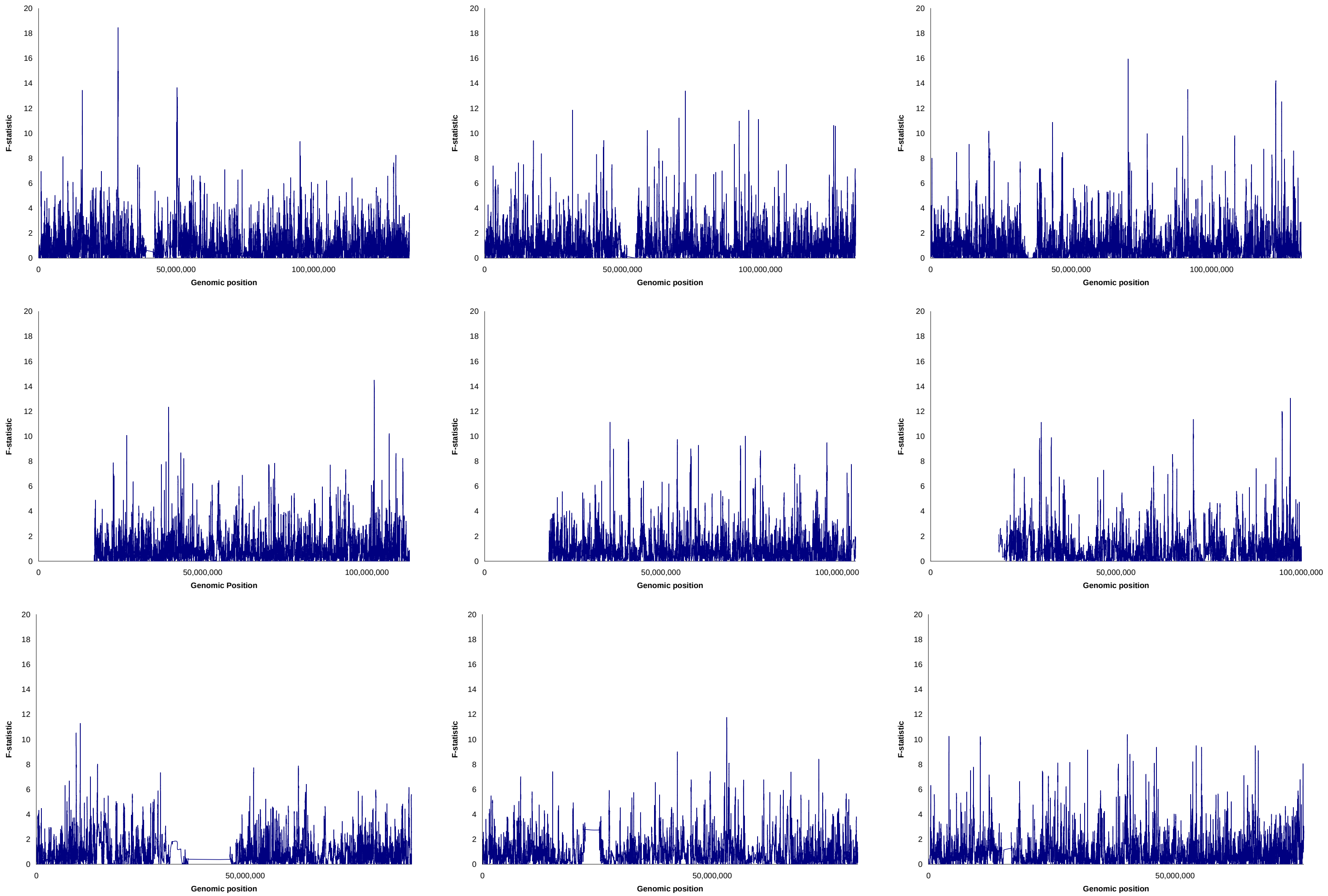


Figure 3. Continued.

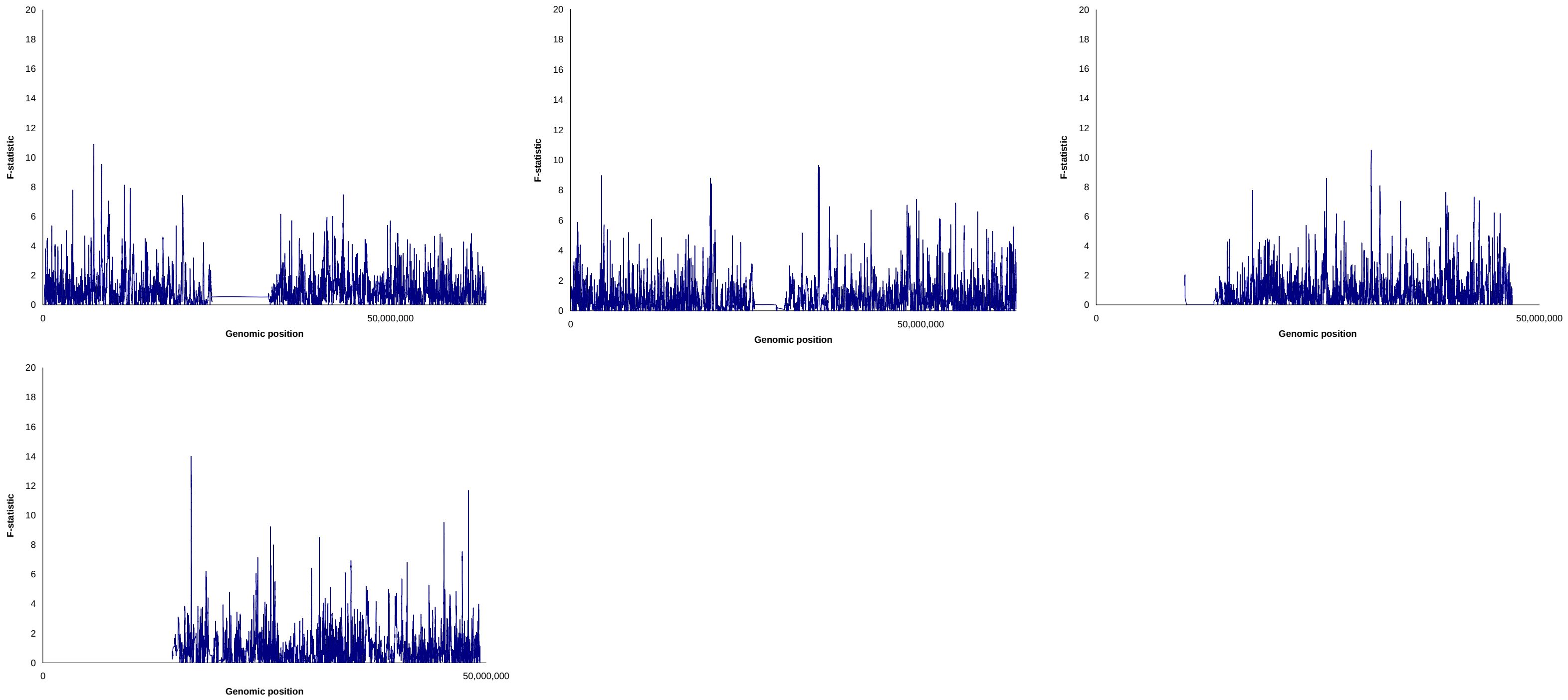


Figure 3. Continued.

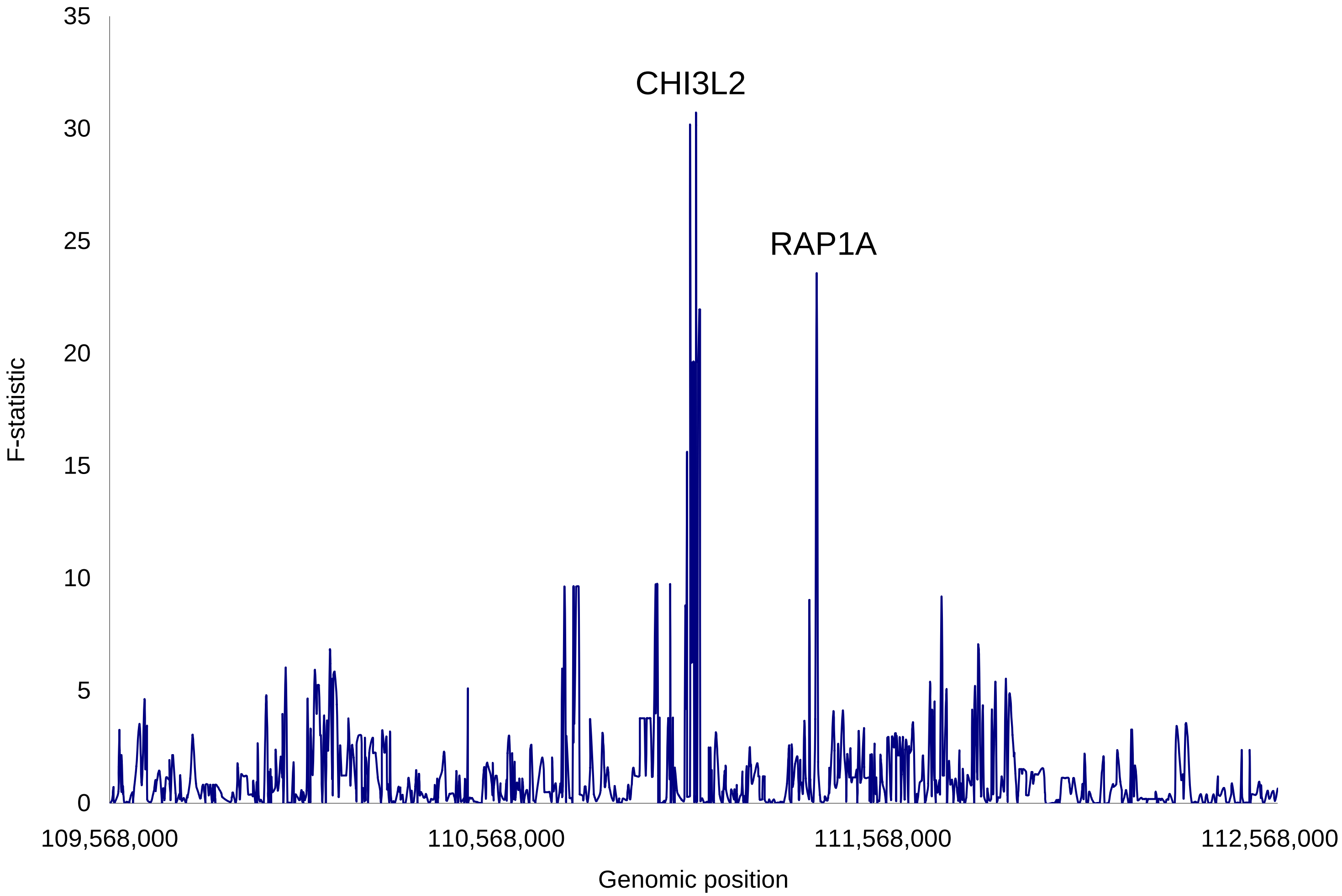


Figure 4. Fine mapping results 3MB around the CHI3L2 region.

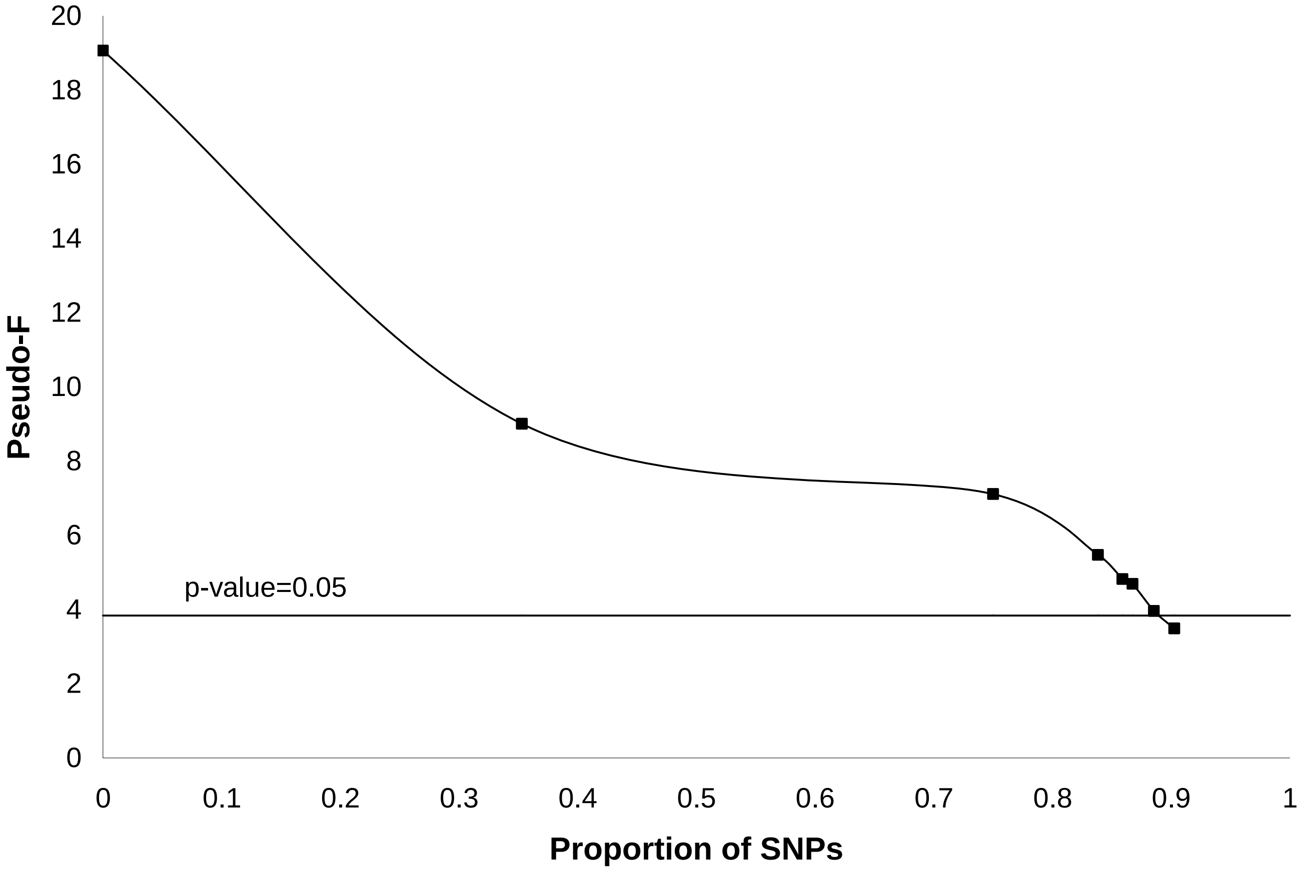


Figure 5. The effect of including known non-associated SNPs in the construction of the haplotype similarity matrix. Essentially, the 11 SNPs in the CHI3L2 gene were initially used to construct the similarity matrix. Additional SNPs from non-associated regions through out the genome were added to those used to construct the similarity matrix in greater numbers. These matrices were then analyzed for association. The Figure suggests that the original signal provided by the 11 CHI3L2 SNPs was so strong that additional SNPs, comprising almost 90% of the SNPs used to construct the matrix, could not completely eliminate the statistical significance of the association.

ACKNOWLEDGEMENTS

The text of Chapter 3 has been submitted for publication as:

> Jennifer Wessel, Ondrej Libiger, and Nicholas J. Schork. Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis.

The dissertation author was the primary researcher and/or author and the co-authors listed in this publication directed and supervised the research which forms the basis for this chapter.

CHAPTER 4

Accommodating Pathway Information in Expression Quantitative Trait Locus (“eQTL”) Analysis

ABSTRACT

The availability of high-throughput genotyping technologies and microarray assays has allowed researchers to consider pursuing investigations whose ultimate goal is the identification of genetic variations that influence the levels at which genes are expressed, e.g., “expression Quantitative Trait Loci” or “eQTL” mapping studies. However, the large number of genes whose expression levels can be tested for association with genetic variations can create both statistical and biological interpretive problems. We consider the integrated analysis of eQTL mapping data that incorporates pathway, functions, and disease process information. The goal of this analysis is to determine if compelling patterns emerge from the data that are consistent with the notion that perturbations in the physiologic environment induced by genetic variations implicate the expression patterns of multiple genes via genetic network relationships or feedback mechanisms. We apply available genetic network and pathway analysis software, as well as a novel regression analysis technique to carry out the proposed studies. We also consider extensions of the proposed strategies and areas of future research.

INTRODUCTION

The exploration of the immediate molecular-physiologic consequences of DNA sequence variation has been greatly enhanced as a result of the introduction of high-throughout, multiplex technologies such as gene expression microarrays, proteomics technologies, and metabolomic assays. A number of studies have been pursued recently that have shown that naturally occurring DNA sequence variations in a wide variety of organisms influence the levels of the expression of particular genes [1-6]. This is not surprising given that DNA sequence variations, such as single nucleotide polymorphisms or deletions, in gene regulatory regions of the genome, such as promoters, could, e.g., influence the ability of a transcription factor to bind and thereby affect the activity of the promoter in guiding transcription of the gene.

Many studies examining the relationship of DNA sequence variations and gene expression levels have not actually considered the biological mechanisms behind such relationships, but have rather focused on the mere association between sequence variations and gene expression patterns in an effort to make broad claims about the role of likely *cis*-regulatory vs. *trans*-regulatory factors in mediating gene expression on a genome-wide scale. These studies, known as expression quantitative trait locus ("eQTL") mapping studies or "genetical genomic" studies, have shed enormous light on the global role of sequence variation in mediating gene expression [1-8]. However, the mere association of a particular genetic variation with the expression level of a gene – whether or not that sequence variation resides within the gene whose expression level is influenced – ultimately raises a number of questions about the relationships of the associations themselves. For example, one could ask if the genes whose expression levels are influenced by a particular genetic variation appear to be involved in the same genetic network, process, or pathway. Addressing such questions could lead to the characterization of genetic variations that influence entire processes and raise the possibility that one of the genes that are influenced by the sequence variation in question is more upstream in the network or process of relevance. Thus, one could infer that a perturbation in a particular gene can induce a cascade of physiologic events that affects all, or many, of the other genes in a particular network or process.

The reason why this type of analysis is important is obvious: it is very unlikely that the expression level of a single gene, when perturbed by a single naturally occurring DNA sequence variation, will induce an overt clinically-identifiable or physiologically meaningful phenotype, given the fact that genes operate in networks replete with redundancy, feedback, and compensatory mechanisms. In fact, it is well known that most traits or diseases are multifactorial and complex genetically, whereby many genes and/or environmental factors are responsible for their expression.

We have therefore considered the analysis of published eQTL mapping studies involving humans that takes into account the possible participation of genes in various networks, pathways, diseases, or drug targets, whose expression values appear to be influenced by particular SNPs. The goal of the analysis is to determine if it is possible to make sense of the collection of genes whose expression patterns are influenced by a group of SNPs. In this light we address two related questions: 1. are the genes whose expression patterns appear to be associated with a particular SNP involved in a particular known process or network? and 2. do the genes whose expression patterns appear to be associated with different SNPs have any commonalities? Or rather, do some sets of SNPs (either working in *cis* or *trans*-acting fashions) influence the expression levels of genes in the same pathway or network?

Although a number of studies have been undertaken in humans to identify genetic variations influencing the expression levels of genes [1-3, 9, 10], we have concentrated on the analysis of 28 SNPs and expression data on 8,523 genes obtained by Cheung et al. (2005) due to its recognition by the scientific community, the availability of the data, and the fact that we have considered these data in particular candidate gene analyses [11] as well as genome-wide association studies 12. To carry out the analysis, we took advantage of Ingenuity's Pathway Analysis Software [13] as well as a novel multivariate analysis technique that can be termed 'multivariate distance matrix regression' (MDMR) analysis that has been shown to have utility in the analysis of high-dimensional gene expression and SNP data [11, 14, 15].

METHODS

CEPH Gene Expression and SNP Data

We used data from Cheung et al. [1, 16] which included gene expression data representing 8,523 unique genes on 57 CEPH-repository-derived individuals whose DNA was studied for polymorphism as part of the International HapMap Project [17]. We note that we actually examined the expression levels of 8672 probes, but some of these probes interrogated the same gene. The HapMap research produced approximately 1.5 million genotypes on these 57 individuals. Cheung et al. (2005) identified strong associations between particular SNP variations and the expression values of particular genes for 24 SNPs (referred to here as 'associated' SNPs) as described in Table 1 of Cheung et al. (2005). In addition to these 24 SNPs, we included in our analyses 4 randomly chosen SNPs as controls, since these SNPs were not found to have strong associations with the expression values of any genes. The additional SNPs had RS numbers of rs10017431, rs10498658, rs2688692, and rs2587021.

Univariate Gene Expression Analysis

Analysis of the relationship of the expression level of each gene to each SNP was performed by testing the equality of expression levels across genotype categories using univariate analyses that included traditional t-tests and the non-parametric Mann Whitney U-test [18]. Some of the 24 SNPs had very unequal allele frequencies, and hence had no or only a few individuals with a particular homozygous genotype. Due to this fact, the rarer homozygote and the heterozygote were combined creating two genotypic categories that could be contrasted for gene expression differences.

Power Studies and False Discovery Rates

We used the techniques described in the paper and website associated with PowerAtlas [19-21], which describes tools designed for microarray data analysis power calculations. The techniques described in the paper and implemented on the website were used to estimate the probability of true positive results (PTP), the probability of true negative results (PTN) and the expected discovery rate (EDR) of the 8523 gene expression association studies across genotype categories for the 28 SNPs. The PTP is defined as the proportion of genes that are declared significantly differentially expressed between two groups, a concept similar to the false discovery rate. The EDR is the average power for all genes for which the null hypothesis is false in an experiment, or in other words the proportion of genes that are differentially expressed that will be declared as such. For each of the 28 SNPs, we report the estimated PTP and EDR at a significance threshold of 0.05 and 0.01.

Pathway Analysis

We used the Ingenuity pathway analysis software package [13] to analyze and assess sets of functions (Gene Ontology (GO) terms), canonical pathways, diseases, and drug targets overrepresented in the lists of genes whose expression levels were influenced by each SNP. We also considered similar analyses that tested for overrepresentation of functions, canonical pathways, and diseases, among genes that were common to a set of SNPs ranked by p-value from Ingenuity. We used the right-tailed Fisher’s exact test, as implemented in

Ingenuity software, that assesses the number of genes in a particular list that participate in a given pathway, relative to the total number of occurrences of these genes in all pathways annotated by Ingenuity. In this way we could determine if sets of SNPs appear to influence, or contribute to, the functioning of particular genetic networks that may be associated with particular phenotypes. We identified the drug targets in the most common network for each SNP. Since our purpose was to capture multiple genes in a pathway affected by a given SNP, we used a less conservative cut-off (p<.05) to increase the probability we would capture the majority of effects exhibited, which could include multiple weak effects.

Multivariate Distance Matrix Regression Analysis

We took advantage of a recently introduced multivariate analysis procedure developed by the authors [11, 14, 15]. This technique, which can be termed 'multivariate distance matrix regression' or 'MDMR' analysis, involves the construction of a distance matrix over the expression values of many genes for the subjects in a study. The similarities and dissimilarities among the subjects based on their gene expression profiles are then related to additional factors collected on those subjects, such as SNP genotype information. The analysis functions in an analogous manner to regression analysis in that the goal is to determine the significance of the additional factors (e.g., SNPs) in 'explaining' the variation in the similarity or dissimilarity of the subjects (based on expression profiles) represented in the distance matrix. We applied this analysis by constructing Euclidean distance matrices for the subjects based on their expression profiles involving genes known to participate in particular processes or pathways and then testing to see if particular SNPs influenced the variation in the similarity/dissimilarity of the gene expression profiles. In effect, we could test the hypothesis that particular SNPs influence the expression profile of genes in an entire pathway or process.

RESULTS

Individual Gene Results

Univariate association analyses involved each SNP we chose to study and the available gene expression data resulting in 28 x 8,523 = 238,644 analysis results (the results of these analyses are available from the authors). We ranked the p-values associated with the analysis of each gene expression variable for each SNP. The genes whose expression values were most strongly associated with each SNP were then used in the pathway analysis. We used an arbitrary p-value cutoff of $p<0.05$ to identify genes whose expression values were "associated" with each SNP. We evaluated the utility of this $p<0.05$ criterion via power and expected discovery rate calculations, as described below. Table 1 provides the results of the analyses for each individual SNP using the Ingenuity software package as described in the Methods section. Table 1 only includes, for the sake of space limitations, the top ten diseases, functions and canonical pathways (ranked by p-values) that were overrepresented in the genes whose expression values were associated ($p<0.05$) with each SNP. The drug targets listed were those represented in the top scoring network for each SNP. The abbreviations used in Table 1 are listed in Appendix 1.

Power Studies

We determined the expected discovery rates (EDR), the probability of a true positive (PTP) result, and the probability of a false positive (PTN) result based on the 8,523 univariate gene expression analyses performed for each SNP using the strategies described by Page et al[19] and Gadbury et al[20], as implemented on the ‘PowerAtlas’ website[21]. This analysis provided us with insights as to the strength of the associations between the SNPs Cheung et al. (2005) identified as significantly associated with the expression levels of a particular gene and the expression levels of other genes. Table 2 provides the results of the analyses as well as the percentage of genes whose expression levels were significantly associated (p<.05 and p<.01) with the SNPs. Most of the 24 SNPs Cheung et al (2005) identified as highly associated with the expression level of a particular gene showed promising EDR or PTP estimates. The most consistently high PTP (≥0.8) and EDR (≥.4) observed were found for SNPs with designations rs10490570, rs10509846, rs4755741 and rs9600337. Many of the other ‘associated’ SNPs described by Cheung et al. (2005) had PTP (≥0.8) (e.g. rs10807387, rs7802273, rs80092794) or EDR (≥.4) (rs3757791, rs788350) with moderately high values for the PTP and/or EDR. We note that a few of Cheung et al’s (2005) ‘associated’ SNPs had low PTP and EDR values (e.g. rs2271194 and rs227940) suggesting that either the associations those

SNPs had with gene expression values were literally confined to one gene, or the results were likely to be false positives. Note the 4 control SNPs (i.e., SNPs not identified by Cheung et al (2005) as having any strong association with any gene's expression values) showed either an EDR or PTP = 0. Figure 1 provides a graphical display of the PTP, EDR, and PTN as a function of sample size based on one of Cheung et al's (2005) associated SNPs (rs10490570) and one control SNP (rs10017431).

Individual SNP Pathway Analysis

We report up to ten of the significantly overrepresented functions, diseases, canonical pathways, and drug targets for each of the 28 SNPs in Table 1. There are 611 unique 'networks' involved in canonical pathways, functions or disease processes queried in the Ingenuity software. For the functions and diseases, there were usually more than ten significantly overrepresented networks for the SNPs, but for the canonical pathways 14.3% of the SNPs had less than ten that were significantly overrepresented. The most significantly overrepresented networks for the SNPs were associated with diseases or particular biological functions (p=0.00002), and these were associated with SNPs rs2139512, rs2762 and rs6928482. We determined if these overrepresented networks included those expression levels that were identified as the most strongly associated with a SNP. As an example, we considered genes whose expression levels were associated with rs6928482. For the 628 genes associated with rs6928482, based on univariate statistical tests, the Ingenuity database identified six genes involved in cellular hematological disorders (univariate p's= 0.003 – 0.04), and the most significantly associated gene, *HAMP*, was in the top 1% of the most strongly associated gene expression levels. Conversely, eight genes were the most significantly associated (p's<.00001) in univariate tests of rs6928482 out of the 8523 expression levels. These eight genes were different from those involved in the cellular hematological disorders, and four of these genes were not involved in any of the functions or diseases identified, while the other four were not involved in any of the top 10 functions or diseases.

Combined SNP Analysis

We identified two pathways that were the most frequently represented among the 28 SNPs we analyzed. These were the Wnt/Beta-catenin signaling pathway (common to 4 of the 28 SNPs) and the serotonin receptor (SR) signaling pathway (also common to 4 SNPs). The Wnt/Beta-catenin signaling pathway was the most significantly associated canonical pathway for three of the SNPs. Of the genes in the Wnt/Beta-catenin signaling pathway, between 8–25 of them were significantly associated with the 4 SNPs (p's<0.05) for which this pathway was overrepresented. Some of these genes overlapped across these 4 SNPs while some were unique to a particular SNP. We then identified 50 Wnt/Beta-catenin signaling pathway genes whose expression levels were influenced by at least one of these four SNPs and used them to form a Euclidean-distance/similarity matrix for the MDMR analysis (see Methods). The MDMR analysis suggested that one of the four SNPs was strongly associated with expression levels of these

50 genes when these genes were considered as providing multivariate gene expression profiles (i.e., when they were considered jointly and not as single gene expression levels assessed individually; Table 3; p=0.0008). This SNP explained 5.0% of the variation in the similarity/dissimilarity of the expression profiles of these 50 genes across the subjects (Table 3). In individual univariate analyses involving genotype categories for this SNP (i.e., rs10807387) and each of the 50 gene expression levels, the p-values ranged from 0.0007 – 0.99, with the expression levels of the WNT2 gene being the most significantly associated with this SNP.

When the similarity analysis was constructed with only those genes whose expression values were significantly associated with an individual SNP based on univariate analyses, the results improved (only significant results shown, Table 4). Two SNPs were actually found to be associated with the similarity of the expression profiles of these genes (Table 4), however only one gene (*LRP5*) was associated with both SNPs from univariate analyses. For the SR signaling pathway, 3 SNPs were identified as being associated with this pathway and involved the expression levels of 4–8 genes (14 in total). Two SNPs were associated with the expression levels of the 14 genes involved in the SR signaling in univariate and multivariate tests, and explained 8.8% of the variation (Table 3).

DISCUSSION

We have shown that a comprehensive, pathway-oriented analysis of eQTL mapping data can lead to more compelling insights about the relationships of DNA sequence variations and gene expression levels. It is well-known that genes participate in networks and do not function in isolation. It is therefore important to consider this fact when evaluating the ultimate significance of the impact of sequence variation on the expression levels of multiple genes. We find that many of the SNPs shown to be strongly associated with the expression levels of particular genes by Cheung et al. [1] are also associated with the expression levels of many other genes, and that the reason the SNP might be associated with these other genes' expression levels is due to the fact that those genes participate in common sets of biological processes or pathways.

We also find that many SNPs, some of which were identified as significantly associated with particular gene expression levels by Cheung et al. 2005, do not appear to be associated with the expression levels of genes participating in common pathways. This suggests that either the SNP affects the expression level of a gene whose influence is non-essential for a particular process (i.e., it is compensated for easily or is peripheral to the activities of that process) or the original association is likely to be a false positive result.

Our assessment of the likely 'expected discovery rates' and 'probability of a true positive' result for each SNP also sheds light on the utility of pathway analyses in eQTL analysis. We find that many of the SNPs found to be significantly associated with the level of expression of a particular gene by

Cheung et al. (2005) did not yield strong associations with other genes and hence had very low expected discovery rates and probabilities of true positive results. Examples include SNPs rs2271194 and rs2762 which Cheung et al. (2005) found to be highly associated with the expression levels of the RPS26 ($p=7.94x10^{-12}$) and LRAP ($p=1.98x10^{-19}$) genes, respectively, but had very low EDR and PTP values (EDR=0.01 for both and PTP=0.02 and 0.18, respectively). This analysis calls into question the actual influence of these SNPs on gene expression patterns despite their strong association with the expression levels of a single gene. We want to emphasize that our use of a common p-value (i.e., $p<0.05$) for declaring significance of an association between each SNP and the 8,523 gene expression levels may not have been ideal. Rather, it might be advantageous in analyses like the one we have pursued to use different thresholds for each gene based on, e.g., EDR and PTP analysis, to generate the list of genes to be interrogated in the pathway and common process analysis. Our analysis of the influence of a SNP on an entire pathway via the MDMR analysis also shows promise in this regard, in that we were able to show that particular SNPs appeared to influence the entire expression profile of a group of genes known (or at least likely) to participate in a particular process or pathway.

Our analysis of processes, diseases, and pathways demonstrate the heterogeneity of functions and diseases that a polymorphic locus can influence. We did find, however, some loci that appear to influence a limited number of networks, for example rs6060535, which may be tightly regulated pathways or pathways with fewer genes involved. The functions we considered generally produced smaller p-values than the disease categories, possibly due to the number of genes involved in either. This could reflect limitations in the Ingenuity knowledgebase regarding disease processes.

There are a number of important issues and directions for future research that should receive special attention. For example, our analysis is entirely dependent on the veracity and/or completeness of the available knowledge-base(s) of biochemical processes and gene interactions (in our case, the knowledge-base developed by Ingenuity). In addition, it is unclear if our analysis of expression patterns observed in immortalized lymphocytes truly correlates with activities and functions *in vivo*, as it is well-known that the expression levels of genes are tissue specific and possibly influenced by transformation [22, 23]. Despite these caveats, our analysis has promise. It may be possible that in the future, as more and more eQTL mapping studies are pursued and knowledge of the interaction of genes grows, one could work out – using purely computational methods – the regulatory machinery associated with not just individual genes, but entire networks, pathways, or processes, mechanisms likely involved in pathway events.

APPENDIX

Abbreviations of canonical pathways, functions, diseases and drug targets used in Table 1.

Description	Abbreviation
alanine and aspartate metabolism	AAM
accumulation of adipose tissue	AAT
activation of Atf-1 binding site	AAtf1BS
adhesion of blastomeres	AB
attachment of breast cancer cell lines	ABCCL
aggregation of B lymphocytes	ABL
adhesion of cell-associated matrix	ACAM
adhesion of cumulus cells	ACC
autophagy of colorectal cancer cell lines	ACCCL
arrest in cell cycle progression of breast cell lines	ACCPCL
arrest in cell cycle progression of lung cell lines	ACCPLCL
arrest in cell cycle progression of brain CA cell lines	ACPBCL
arrest in development of cells	ADC
attention deficit hyperactivity disorder	ADHD
arrest in differentiation of leukocytes	ADL
adhesion of ova	AdO
autoimmune disease of primate	ADP
ADP ribosylation of amino acids	ADPRAA
arrest in development of thymocytes	ADT
arrest in development of T lymphocytes	ADTL
adhesion of embryonic cells	AEC
apoptosis of ectodermal cells	AECC
anoikis of epithelial cell lines	AECL
aggressive fibromatosis	AF
arrest in G1 phase of thyroid tumor cell lines	AG1PTTCL
aggregation of breast cancer cell lines	AgBCCL
aggregation of epithelial cells	AgEC
aggregation of tumor cell lines	AgTCL
activation of HNF4 binding site	AHNF4
apoptosis of intestinal cell lines	AICL
arrest in interphase of skin cancer cell lines	AISCCL
arthritis of joint	AJ
attachment of kidney cells	AKC
accumulation of lung cancer cell lines	ALCCL
acidification of leukemia cell lines	ALCL
allodynia of rodents	AlR
ataxia of mice	AM
adhesion of muscle cells	AMC

attachment of microtubules	AMt
activation of nerves	AN
apoptosis of neuroblasts	ANB
apoptosis of natural killer T lymphocytes	ANKTL
angiogenesis of tumor	AnT
atherogenesis of organism	AO
abdominal obesity-metabolic syndrome	AOMS
amyloid Processing	AP
acute pancreatitis	APC
adhesion of pancreatic cancer cell lines	APCCL
apoptosis of endothelial cells	ApEC
adiposis of mice	ApM
apoptosis of myoblasts	ApMb
apoptosis of mast cells	ApMC
activation of plasma membrane projections	APMP
apoptosis of osteoblasts	ApO
assembly of protein-protein complex	APPC
apoptosis of splenocytes	ApSc
apoptosis of thymoma cell lines	ApTCL
activation of permeability transition pores	APTP
anorexia of rodents	AR
anisocytosis of red blood cells	ARBS
antiviral response of melanoma cell lines	ARMCL
accumulation of RNA	ARNA
arrhythmogenic right ventricular dysplasia	ARVD
Alport's syndrome	AS
aggregation of squamous cell carcinoma cell lines	ASCCCL
aggregation of stomach cancer cell lines	ASCCL
activation of steroidogenic factor-1 response element	ASF1RE
apoptosis of spiral ganglion cells	ASGC
aminosugars metabolism	ASM
apoptosis of skeletal muscle cells	ASMC
activation of STAT response element	ASTATRE
astrocytosis of tissue	AT
accumulation of tumor cell lines	ATCL
angiogenesis of tissue	ATI
activation of T lymphocytes	ATL
arthritis of rats	AtR
aminoacyl-tRNA biosynthesis	AtRNAB
attachment of tumor cell lines	AtTCL
atrial ventricular block	AVB
accumulation of white adipose tissue	AWAT
anxiety of mice	AxM
anxiety of rodents	AxR
bronchial asthma	BA

binding of adenocarcinoma cells	BAC
beta-alanine metabolism	BAM
bile acid biosynthesis	BAS
binding of AT rich element	BATRE
binding of bone cell lines	BBCL
branching of breast cell lines	BBrCL
basal-cell carcinoma	BCC
binding of CD28RE/AP response element	BCD28R
B cell receptor signaling	BCRS
bundling of filaments	BF
blood group glycolipid biosynthesis-neolactoseries	BGGBN
binding of hormone	BH
binding of leukocyte cell lines	BLCL
blistering of mice	BM
binding of mRNA	BmRNA
bundling of microtubules	BMt
beta-oxidation of lignoceric acid	BOLA
biosynthesis of protein	BP
binding of T lymphocytes	BTL
butanoate metabolism	ButM
chylomicronemia	C
conversion of adenosine	CA
cerebral amyloid angiopathy of mice	CAAM
contraction of actin cytoskeleton	CAC
chemoattraction of monocytes	CaMo
cAMP-mediated signaling	cAMPMS
cancer signaling	CaS
coagulation of blood	CB
cardiac beta-adrenergic signaling	CBAS
coagulation of bodily fluid	CBF
contraction of blood vessel	CBV
citrate cycle	CC
chemotaxis of cancer cells	CCaC
complement and coagulation cascades	CCC
cell cycle: G1/S checkpoint regulation	CCG1/S
cell cycle: G2/M DNA damage checkpoint regulation	CCG2/M
cardiac contractility of heart	CCH
communication of cell lines	CCL
cell cycle progression of breast cell lines	CCPBCL
cell cycle progression of muscle cell lines	CCPMCL
cell division of germ cells	CDGC
cell death of lung cell lines	CDLCL
cleavage of DNA	CDNA
contact dermatitis of organ	CDO
cell death of spinal cord cells	CDSCC

conversion of embryonic cells	CEC
communication of endothelial cell lines	CECL
chemotaxis of eukaryotic cells	CEkC
colony formation of colony forming unit-megakaryocytes	CFCFUM
cytolysis of fibroblast cell lines	CFCL
colony formation of pre-B lymphocytes	CFPBL
contact growth inhibition of breast cell lines	CGIBCL
contact growth inhibition of colon cancer cell lines	CGICCCL
contact growth inhibition of cervical cancer cell lines	CGICvC
contact growth inhibition of epithelial cell lines	CGIECL
communication of gap junctions	CGJ
congenital heart block	CHB
cytostasis of hepatoma cell lines	CHCL
chemokinesis of cell lines	CkCL
compartmentalization of leukocytes	CL
cytostasis of lung cancer cell lines	CLCCL
chemotaxis of lymphoma cell lines	CLCL
cross-linkage of DNA	CLDNA
chemotaxis of leukemia cell lines	CLkCL
convulsion of mice	CM
chemotaxis of memory B lymphocytes	CMBL
chemotaxis of monocytes	CMc
cell movement of carcinoma cells	CMCC
cancer of mammary gland	CMG
cell movement of germ cells	CMGC
cell movement of kidney cell lines	CMKCL
cell movement of lymphoma cell lines	CMLCL
cell movement of mammary tumor cells	CMMTC
chemotaxis of macrophages	CMp
cell movement of peripheral blood leukocytes	CMPBL
capping of mRNA	CmRNA
chemotaxis of memory T lymphocytes	CMTL
chemotaxis of natural killer cells	CNKC
chemotaxis of natural killer T lymphocytes	CNKTL
communication of nervous tissue cell lines	CNTCL
chemotaxis of peripheral blood monocytes	CPBM
cognition of rodents	CR
cytosis of red blood cells	CRBC
cell rounding of pheochromocytoma cell lines	CRPCL
chemokine signaling	CS
contraction of stellate cells	CSC
compaction of stomach cancer cell lines	CSCCL
congenital stationary night blindness, type 1	CSNB1
chemotaxis	Ct
chemotaxis of Th2 lymphocytes	CTh2L

clustering of telomeres	CTm
cell viability of colon cancer cell lines	CVCCCL
cell viability of colorectal cancer cell lines	CVCRCCL
cardiovascular disorder	CVD
cell viability of embryonic stem cells	CVESC
cysteine metabolism	CysM
dysmyelination of axons	DA
development of adenocarcinoma	Dac
deposition of amyloid fibrils	DAF
delay in apoptosis of neurons	DAN
depletion of ATP	DATP
disruption of blood-brain barrier	DBBB
degranulation of bone cell lines	DBCL
diabetes of mice	DbM
differentiation of bone-marrow-derived monocyte/macrophage precursor cells	DBMD
damage of colon	DC
depletion of Ca	DCa
differentiation of carcinoma cell lines	DCCL
damage of cardiovascular tissue	DCVT
development of digit	DD
delay in differentiation of epithelial cells	DDEP
deval disorder of gonad	DDG
deval disorder of heart ventricle	DDHV
delay in development of lymphatic system cells	DDLSC
deval disorder of mammary duct	DDMD
delay in differentiation of sensory epithelium	DDSE
dysgenesis of eye	DE
deformation of mandible	DfM
depolarization of ganglion cells	DGC
D-glutamine and D-glutamate metabolism	DGDGM
development of genitourinary tract	DGT
development of head	DH
damage of heart cell lines	DHCL
development of hypothalamic nucleus	DHN
development of heart tube	DHT
differentiation of myelomonocytic cells	DMC
density of microtubules	DMt
degeneration of neurites	DN
differentiation of neuroblastoma cell lines	DNCL
delay in neurological disorder of mammalia	DNDM
dysgenesis of outflow pathway	DOP
depolarization of cells	DpC
development of prostate gland	DPG
detachment of retina	DR

development of renal glomerulus	DRG
differentiation of spermatids	DS
differentiation of Sertoli cells	DSC
depolarization of superior cervical ganglion neurons	DSCGN
development of skeletomuscular system	DSS
damage of seminiferous tubules	DST
damage of tumor cell lines	DTCL
damage of tumor cell lines	DTCL
development of skin	DvS
edema	E
ERK/MAPK signaling	E/M
extension of actin cytoskeleton	EAC
experimental allergic encephalomyelitis (chronic relapsing) of mice	EAEM
epidermolysis bullosa	EB
epidermolysis bullosa simplex	EBS
exposure of Ca2+	ECa
engulfment of cell lines	ECL
expression of cytokine response element	ECRE
expansion of dendritic cells	EDC
epidermal hyperplasia	EdH
enlargement of endosomes	EE
erythropoiesis of embryonic stem cells	EESC
EGF signaling	EGFS
endocrine gland tumor	EGT
exocytosis of histamine	EH
ejaculation	Ej
ejaculation of mice	EjM
epilepsy	El
exudation of lung pleura	ELP
enchondromatosis	EM
endometrial hyperplasia	EmH
elongation of mRNA	EmRNA
ectopia of neurons	EN
early-onset morbid obesity	EOMO
erythropoietic protoporphyria	EP
ER signaling	ERS
endoplasmic reticulum stress pathway	ERSP
eicosanoid signaling	ES
engulfment of tumor cell lines	ETCL
fibrosis	F
farnesylation of amino acids	FAA
fatty acid biosynthesis (Path 1)	FAB1
fatty acid biosynthesis (Path 2)	FAB2
fatty acid metabolism	FAM

formation of carcinoma	FC
function of liver	FcL
formation of T-cell non-Hodgkin lymphoma	FCNHL
function of tissue	FcT
fibrosis of dermis	FD
formation of elastic fibers	FEF
FGF signaling	FGFS
formation of hepatocellular carcinoma	FHC
familial hemiplegic migraines	FHM
fibrosis of interstitial tissue	FIT
fibrosis of liver	FL
fibrillogenesis of tissue	FlT
fibrosis of organ	FO
folding of protein	FP
formation of pseudopodia	FPp
formation of superoxide radical	FSR
fibrosis of tissue	FT
fusion of tumor cell lines	FTCL
formation of tectorial membrane	FTM
fusion of lysosome	FuL
fusion of phagosomes	FuP
flux of chlorine	FxCl
glycogenesis	G
G0/S phase transition of cell lines	G0/SPT
G2/M phase of colorectal cancer cell lines	G2/MP
growth of adenoma	GA
GABA receptor signaling	GABARS
generalized atrophic benign epidermolysis bullosa	GABEB
ganglioside biosynthesis	GB
growth of breast carcinoma	GBC
growth of carcinoma	GC
glycosaminoglycan degradation	GD
growth of lung cancer cell lines	GLCCL
glutamate metabolism	GluM
growth of mast cells	GMC
GM-CSF signaling	GMCSFS
growth of malignant tumor	GMT
glomerulonephritis of mice	GnM
growth of ovarian cells	GOC
growth of papilloma	GP
GPCR signaling	GPCRS
glycerophospholipid metabolism	GPM
G-protein signaling, coupled to cyclic nucleotide second messenger	GPSC
growth of primary tumor	GPT

glutamate receptor signaling	GRS
glomerulosclerosis	GS
glomerulosclerosis of mice	GsM
glycine, serine and threonine metabolism	GSTM
growth of ureteric bud cells	GUBC
hemangioma	H
hyperalgesia	Ha
hemolytic anemia of mice	HAM
hemorrhage of brain	HB
hyperproliferation of breast cell lines	HBCL
hyperplasia of bone marrow cells	HBMC
hypercholesterolemia	Hc
hemorrhage of cerebrum	HCb
hematologic cancer of humans	HCH
homing of cell lines	HCL
hematological disorder of cells	HDC
hematological disorder of eukaryotic cells	HDEC
hematological disorder of heart	HDH
hematological disorder of rats	HDR
homing of eukaryotic cells	HEC
hyperplasia of exocrine gland	HEG
heart failure	HF
hyperplasia of follicular cells	HFC
hypogonadism	Hg
hyperplasia of gonadal cells	HGC
hydrolysis of GDP	HGDP
hypoinsulinemia of mice	HiM
hypertrophy of left ventricle	HLV
hepatomegaly	Hm
hepatitis	Hp
hepatitis C	HpC
hyperproliferation of cell lines	HpCL
hydrolysis of phosphatidylethanolamine	HPd
hypalgesia of rodents	HR
homologous recombination of DNA	HRDNA
hemorrhagic shock	HS
homing of stem cells	HSC
hypoplasia of secretory structure	HSS
hyperthyroidism	Ht
healing of tibia	HTb
hypertriglyceridemia	HTg
hypertriglyceridemia of rodents	HTgR
hypothermia of mice	HtM
hypertension	HTN
induction of B lymphocytes	IBL

intraductal carcinoma	IC
islet-cell carcinoma	ICC
interphase of cervical cancer cell lines	ICCCL
islet cell tumor	ICT
induction of cytotoxic T lymphocytes	ICTL
insulin-dependent diabetes mellitus of mice	IDDMM
interaction of DNA	IDNA
inflammatory disorder of skin	IDS
infiltration of eosinophils	IE
inflammation of eukaryotic cells	IEC
infiltration of fibroblasts	IF
infection of fibroblast cell lines	IFCL
interferon signaling	IFS
infiltration of granulocytes	IG
IGF-1 signaling	IGF1S
infection of human immunodeficiency virus type 1	IHIV1
immortalization of hematopoietic progenitor cells	IHPC
invasion of intestinal cell lines	IICL
inflammation of knee	IK
induction of lymphocytes	IL
IL-10 signaling	IL10S
IL-2 signaling	IL2S
IL-4 signaling	IL4S
IL-6 signaling	IL6S
infection of lymphoma cell lines	ILCL
inositol metabolism	IM
innervation of neurons	IN
inflammation of skin	InfS
inflammation of organ	IO
inflammatory response of mice	IRM
insulin receptor signaling	IRS
integrin signaling	IS
insulitis of mice	IsM
induction of serum response element	ISRE
invasion of tissue	IT
induction of T lymphocytes	ITL
initiation of translation of protein	ITP
invasion of lung cell lines	IvLCL
invasion of T lymphocytes	IvTL
joining of DNA fragment	JDNAF
juvenile rheumatoid arthritis	JRA
kindling	K
killing of fibroblasts	KF
leakage of blood	LB
leakage of blood-brain barrier	LBBB

lysis of blood clot	LBC
lysine degradation	LD
leukemogenesis of humans	LH
lipolysis of lipid	LL
leakage of lysosome	LLs
loss of neurons	LN
loss of oocytes	LO
loss of brain cells	LoBC
learning of rodents	LR
lifespan of T lymphocytes	LTL
memory	M
modification of anion	MA
mobilization of antigen presenting cells	MAPC
modification of adenosine	Mas
modification of chromatin	MC
mobilization of Ca	Mca
modification of chromosome components	MCC
mitosis of carcinoma cell lines	MCCL
morphology of cardiomyocytes	MCm
metabolism of dopamine	MD
macular dystrophy, vitelliform	MDV
morphology of fur	MF
morphology of fibroblasts	MFb
morphology of fibrosarcoma cell lines	MFCL
migration of glioblastoma cells	MGC
morphology of hepatoma cell lines	MHCL
mass of intestine	MI
morphology of liver cells	MLC
morphology of lung cancer cell lines	MLCCL
mitogenesis of leukemia cell lines	MLCL
methane metabolism	MM
metaplasia of mammary gland tissue	MMGT
migration of mammary tumor cells	MMTC
maturation of neurons	MN
modification of polyols	MP
myeloproliferative syndrome of mice	MSM
mitosis of smooth muscle cells	MSMC
morphology of thyroid tumor cell lines	MTTCL
neuroblastoma	N
neurotrophin/Trk signaling	N/TrkS
necrosis of cardiomyocytes	NC
nucleotide excision repair pathway	NERP
N-glycan biosynthesis	NGB
nitrogen metabolism	NgM
natural killer cell signaling	NKCS

neutropenia of mice	NM
necrosis of muscle cells	NMC
nitrogen oxide signaling in the cardiovascular system	NOSCVS
necrosis of parenchymal cells	NPC
neurodegeneration of retinal ganglion cells	NRGC
neuregulin signaling	NS
non-small-cell lung carcinoma	NSCLC
osteolysis of bone	OB
one carbon pool by folate	OCPF
O-glycan biosynthesis	OGB
O-glycosylation of protein	OGP
overflow of norepinephrine	ON
organization of plasma membrane projections	OPMP
osmotic water permeability of oocytes	OWPO
plasmacytosis	P
p38 MAPK signaling	p38MAPKS
pathfinding of axons	PA
phagocytosis of blood cells	PBC
proliferation of bladder cancer cell lines	PBCCL
prostate cancer	PC
pancreatic carcinoma	PcC
proliferation of colon cell lines	PCCL
phagocytosis of cell lines	PCL
prostatic carcinoma	PCn
phagocytosis of phagocytes	PcP
pancreatitis of rodents	PcR
phospholipid degradation	PD
PDGF signaling	PDGFS
polydactyly of limb	PdL
paralysis of hindlimb	PH
phenylalanine metabolism	PheM
proliferation of helper inducer T lymphocytes	PHITL
PI3K/AKT signaling	PI3K
prostatic intraepithelial neoplasia	PIN
prostatic intraepithelial neoplasia of mice	PINM
perturbation of lipid	PL
propanoate metabolism	PM
progressive motor neuropathy of mice	PMNM
polarization of podosomes	PP
PPAR signaling	PPARS
proliferation of pancreatic cancer cell lines	PPCCL
phosphorylation of protein fragment	PPF
paralysis of rats	PR
Parkinson's signaling	PS
proliferation of somatic cells	PSC

primary systemic carnitine deficiency	PSCD
pancreatic tumor	PT
PTEN signaling	PTENS
phenylalanine, tyrosine and tryptophan biosynthesis	PTTB
proliferation of thyroid tumor cell lines	PTTCL
purine metabolism	PurM
pyruvate metabolism	PyrM
quantity of 12-hydroxyeicosatetraenoic acid	Q12HA
quantity of 12(S)-hydroxyeicosatetraenoic acid	Q12SHA
quantity of amino acids	QAA
quantity of beta-estradiol	QBE
quantity of Ca	QC
quantity of Ca2+	QC2
quantity of colony-forming erythroid cells	QCFEC
quantity of cellular inclusion bodies	QCIB
quantity of choline-phospholipid	QCP
quantity of embryonic cells	QEC
quantity of granulocyte-macrophage progenitor cells	QGMPC
quantity of intercellular junctions	QIJ
quantity of inositol phosphate	QIP
quantity of L-triiodothyronine	QLT
quantity of mice	QM
quantity of multilineage progenitor cells	QMPC
quantity of neurites	QN
quantity of ovarian follicle	QOF
quantity of peripheral blood leukocytes	QPBL
quantity of pathological cyst	QPC
quantity of phosphatidylinositol 3,4-diphosphate	QPD
quantity of sarcoma cells	QSC
quantity of trophoblast cells	QTC
release of arachidonic acid	RAA
recruitment of cells	RC
release of cyclic AMP	RCAMP
ruffling of cervical cancer cell lines	RCCCL
repression of cDNA	RcDNA
retraction of cellular protrusions	RCP
remodeling of chromatin	RCt
recruitment of eukaryotic cells	REC
regression of carcinoma	RgC
release of glutamine family amino acid	RGFAA
regulation of tissue	RgT
release of histamine	RH
release of lipid	RL
rolling of lymphoid cells	RLC
retinol metabolism	RM

regression of mullerian duct	RMD
recruitment of osteoclasts	RO
redistribution of phospholipid	RP
release of phosphatidic acid	RPA
replication of smooth muscle cells	RSMC
repression of synthetic promoter	RSP
regression of tissue	RT
ruffling of tumor cell lines	RTCL
release of testosterone	RTt
renal and urological disorder	RUD
renal and urological disorder of mice	RUDM
sedation	S
social anxiety disorder of mice	SADM
SAPK/JNK signaling	SAPK/JNK
sterol biosynthesis	SB
survival of bladder cancer cell lines	SBCCL
secretion of bodily fluid	SBF
stimulation of B lymphocytes	SBL
sprouting of blood vessel	SBV
shortening of cardiomyocytes	SC
shape change of axons	SCA
survival of carcinoma cell lines	SCCL
shape change of connective tissue cells	SCCTC
stilbene, coumarine and lignin biosynthesis	SCLB
small-cell lymphocytic lymphoma	SCLL
shape change of tumor cell lines	SCTCL
synthesis of D-glucose	SDG
synthesis and degradation of ketone bodies	SDKB
stimulation of eosinophils	SE
size of gonadal cells	SGC
sonic hedgehog signaling	SHS
syndactyly of limb	SL
size of late endosomes	SLE
steatohepatitis of mice	SM
stabilization of mouse X chromosome	SMXC
stimulation of neurons	SN
synthesis of protein	SP
sexual receptivity of mice	SRM
serotonin receptor signaling	SRS
synthesis of sterol	SS
survival of stomach cancer cell lines	SSCCL
stimulation of bone marrow cells	StBMC
synaptic transmission of synapse	STS
schizophrenia	Sz
transformation of antigen presenting cells	TAPC

thickness of bone	TB
transport of bicarbonate	TBc
testicular cancer	TC
thickness of carotid artery	TCA
tumorigenesis of carcinoma cells	TCC
transcription of CD28RE/AP response element	TCD28
T cell receptor signaling	TCRS
thickness of connective tissue	TCT
thickening of epithelial tissue	TET
tumorigenesis of fibroblasts	TF
TGF-B signaling	TGFBS
transcription of IL-4 response element	TIL4RE
transformation of lung cancer cell lines	TLCCL
tryptophan metabolism	TM
transport of monocarboxylic acid	TMA
tumorigenesis of ovary	TO
transport of oleic acid	TOA
total peripheral resistance	TPR
transactivation of Runx2 binding site	TRunx2
transactivation of Sf1 binding site	TSf1BS
transcription termination of DNA	TTDNA
transport of vesicles	TV
tyrosine metabolism	TyrM
uremia	U
ubiquinone biosynthesis	UB
unfolded protein response of cells	UPRC
VEGF signaling	VEGFS
VIPoma	VIP
viral life cycle	VLC
valine, leucine and isoleucine degradation	VLID
ventricular tachycardia	VT
ventricular tachycardia of heart	VTH
weight loss of rodents	WLR
wnt/beta-catenin signaling	Wnt/BCS
X-linked mental retardation	XLMR
xenobiotic metabolism signaling	XMS
xeroderma pigmentosum, complementation group E	XPCGE

Drug Name	**Abbreviation**
(6R)-tetrahydrobiopterin	6RT
abatacept	A
amiloride, amiloride/hydrochlorothiazide	AAH
acetaminophen/pentazocine, levorphanol, buprenorphine, naltrexone, pentazocine, naloxone, butorphanol	Ac

adenosine, dyphylline, aminophylline, theophylline, caffeine	ADAT
adalimumab, etanercept, infliximab, CDP870, golimumab, thalidomide	AET
antihemophilic factor, dalteparin, heparin, coagulation Factor VIIa, enoxaparin, coagulation factor IX, rivaroxaban, deligoparin, idraparinux, tifacogin	AFD
alefacept, siplizumab	AS
arofylline, tetomilast, anagrelide, cilomilast, milrinone, roflumilast, caffeine	ATAC
AVP, conivaptan, lypressin	AVP
AVP, lypressin	AVPL
bicalutamide, flutamide, nandrolone decanoate, testosterone cypionate, oxandrolone, danazol, stanozolol, testosterone, oxymetholone, testosterone propionate, testosterone enanthate	BFNT
bevacizumab, pegaptanib	BP
bexarotene, retinoic acid, 9-cis-retinoic acid	BR
clevidipine, amlodipine/benazepril, diltiazem, verapamil, bepridil, enalapril/felodipine, amlodipine/atorvastatin, nisoldipine, isradipine, felodipine, nimodipine, nitrendipine, amlodipine, nicardipine, nifedipine, trandolapril/verapamil, diltiazem/enalapril	CADV
calcitonin (salmon)	CaT
cladribine	CB
collagenase	Cg
cinacalcet	CL
cetrorelix, triptorelin, abarelix	CTA
delta-aminolevulinic acid	DAA
3,5-diiodothyropropionic acid, amiodarone, thyroxine	DAAT
dyphylline, aminophylline, cilostazol, amrinone, theophylline	DACA
darapladib	DP
disulfiram	DS
erythropoietin, darbepoetin alfa	EDA
elsamitrucin, irinotecan, topotecan, rubitecan, gimatecan, karenitecin	EIT
EMD121974	EMD
etoposide, pixantrone, becatecarin, elsamitrucin, AQ4N, mitoxantrone, tirapazamine, nemorubicin, epirubicin, doxorubicin, daunorubicin	EPB

17-alpha-ethinylestradiol, fulvestrant, beta-estradiol, bazedoxifene, ethinyl estradiol/desogestrel, ethinyl estradiol/drospirenone, premarin, ethinyl estradiol/norelgestromin, ethinyl estradiol/norethindrone, ethinyl estradiol/levonorgestrel, ethinyl estradiol/norgestrel, ethinyl estradiol/norgestimate, conjugated estrogen/medroxyprogesterone acetate, FC1271A, toremifene, tamoxifen, raloxifene, arzoxifene, clomiphene, estramustine phosphate, diethylstilbestrol diphosphate	Est
Enzastaurin	EZ
enzastaurin, ruboxistaurin	EZR
5-fluorouracil, AG 337, capecitabine, trifluridine, floxuridine, LY231514	FCTF
Forodesine	FD
Flavopiridol	FP
Gemcitabine	GC
IDN-6556	IDN
interferon gamma-1b	IG1B
IGF1	IGF1
Ipilimumab	IM
isoflurane, mecamylamine, succinylcholine, rocuronium, doxacurium, mivacurium, pipecuronium, rapacuronium, metocurine, atracurium, cisatracurium, acetylcholine, nicotine, D-tubocurarine, enflurane, pancuronium, vecuronium	IMSR
INO-1001	INO
Imatinib	IT
Ketoconazole	KC
LY231514	LY
methazolamide, acetazolamide, dorzolamide, dorzolamide/timolol, brinzolamide	MAD
menotropins, hCG	MH
nelarabine, clofarabine, fludarabine phosphate, cytarabine, trifluridine	NCF
naltrexone, naloxone	NN
Nesiritide	NR
Natalizumab	NZ
Oblimersen	OM
Octreotide	OT
prostaglandin E1	PE1
Propylthiouracil	PT
propylthiouracil, methimazole	PTR
PXD101, vorinostat, FR 901228	PVF
rimonabant, delta-9-tetrahydrocannabinol	RDT

(R)-flurbiprofen	RF
rosiglitazone, GI262570, pioglitazone, tesaglitazar	RGPT
Ruboxistaurin	RS
Riluzole	RZ
sulfasalazine, balsalazide, 5-aminosalicylic acid, verteporfin	SBAV
sunitinib, imatinib, sorafenib, becaplermin	SISB
Sargramostim	SM
SR 48968	SR
saxagliptin, talabostat	ST
TAK-242	TAK
Tiagabine	TG
Thrombin	TH
tranylcypromine, phenelzine, isocarboxazid	TPI
UK-427,857, vicriviroc	UV
XR9576, valspodar	XV
YM 529	YM
Zafirlukast	ZK

REFERENCES

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

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

3. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302

4. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701-3

5. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3: e267

6. Yarovaya N, Schot R, Fodero L, McMahon M, Mahoney A, Williams R, Verbeek E, de Bondt A, Hampson M, van der Spek P, Stubbs A, Masters CL, Verheijen FW, Mancini GM, Venter DJ (2005) Sialin, an anion transporter defective in sialic acid storage diseases, shows highly variable expression in adult mouse brain, and is developmentally regulated. Neurobiol Dis 19: 351-65

7. Rosa GJ, de Leon N, Rosa A (2006) A review of microarray experimental design strategies for genetical genomics studies. Physiol Genomics

8. Kendziorski C, Wang P (2006) A review of statistical methods for expression quantitative trait loci mapping. Mamm Genome 17: 509-17

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

10. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78

11. Wessel J, Schork NJ (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79: 792-806

12. Wessel J, Libiger O, Schork NJ (2006) Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis. Genetic Epidemiology Submitted.

13. Ingenuity's Pathway Analysis Software. http://www.ingenuity.com/.

14. Zapala M, Schork NJ (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. . Proc Natl Acad Sci U S A in press.

15. Ye C, Zapala M, Kang HM, Wessel J, Eskin E, Schork NJ (2006) High-Density QTL Mapping to Identify Phenotypes and Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways. Mol Syst Biol submitted.

16. Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/

17. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437: 1299-320

18. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods.* Pacific Grove: Duxbury Press, 1998.

19. Page GP, Edwards JW, Gadbury GL, Yelisetti P, Wang J, Trivedi P, Allison DB (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics 7: 84

20. Gadbury GL, Page GP, Edwards JW, Kayo T, Prolla TA, Weindruch R, Permana PA, Mountz JD, Allison DB (2004) Power and sample size estimation in high dimensional biology. Statistical Methods in Medical Research 13: 325-38

21. Allison DB, Page GP, Edwards JW. PowerAtlas. http://www.poweratlas.org/.

22. Cotsapas CJ, Williams RB, Pulvers JN, Nott DJ, Chan EK, Cowley MJ, Little PF (2006) Genetic dissection of gene regulation in multiple mouse tissues. Mamm Genome 17: 490-5

23. Shukla SJ, Dolan ME (2005) Use of CEPH and non-CEPH lymphoblast cell lines in pharmacogenetic studies. Pharmacogenomics 6: 303-10

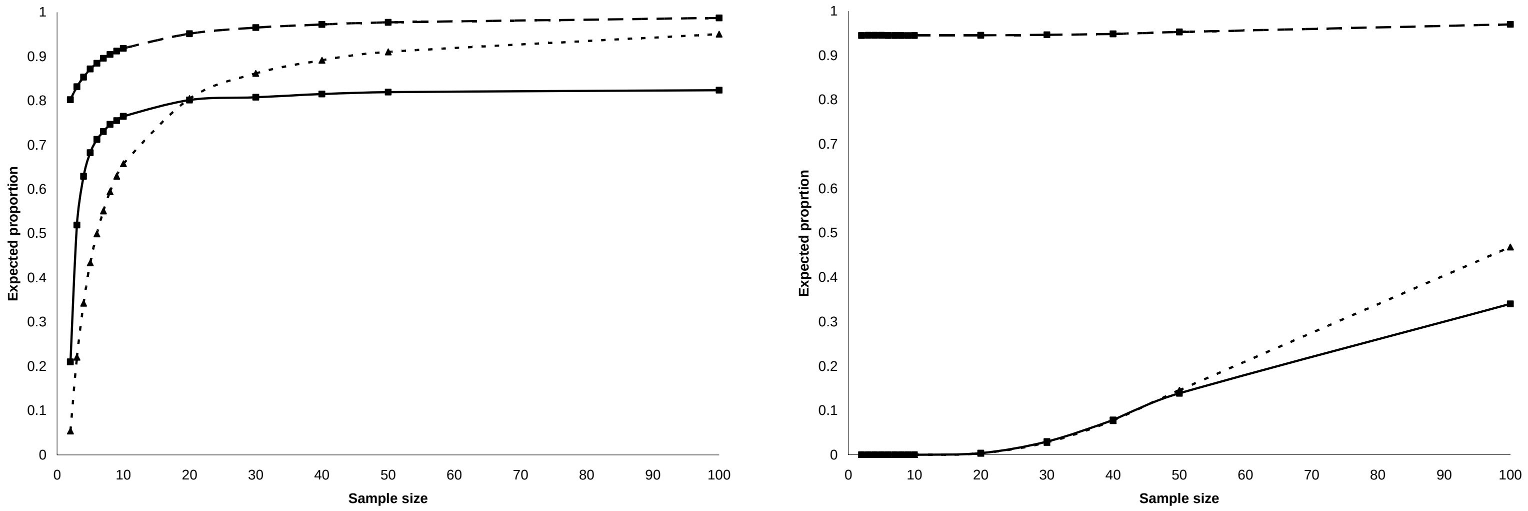


Figure 1. Example power analysis of the SNPs rs10490570 and rs10017431 reporting the true positive, true negative and expected discovery rates as a function of sample size using the methods associated with the PowerAtlas website (www.poweratlas.org). The probability of a true positive (PTP) is reflected by the solid line, the expected discovery rate (EDR) is reflected by the dotted line, and the probability of a true negative result (PTN) is reflected by the dashed line.

Table 1. Gene Expression Analysis Results Assuming Functional, Disease, Pathway, and Drug Target Groupings of the Genes Whose Expression Levels were Significantly Associated with each SNP.

Phenotype	DDX17	VAMP8	CTBP1	ICAP-1A	TM7SF3
SNP rs#	rs10490570	rs10509846	rs1060043	rs10807387	rs11822822
Functions	**ACCPLCL**	**OPMP**	*IG*	**SS**	**FcT**
Diseases	*DMt*	**IL**	*SGC*	*RAA*	**SN**
	PSC	*TIL4RE*	*OPMP*	*HSC*	**ApM**
	SRS	*ITL*	*Mas*	*IRM*	*OGP*
	HtM	*EH*	*RTt*	*FxCl*	*SCCTC*
	MFb	*CBF*	*DATP*	*RL*	*CCPMCL*
	DpC	*EDC*	*IE*	*ECa*	*HGDP*
	CDGC	*CB*	*TV*	*HPd*	*QEC*
	DGC	*Wnt/BCS*	*ACCPLCL*	*SDG*	*LL*
	EmRNA	*IBL*	*RSP*	*CMGC*	*ASTATRE*
	DTCL	**CLCCL**	*CRPCL*	***CVD***	*ANKTL*
	EdH	**CLkCL**	*IsM*	**ABCCL**	*AR*
	DDHV	*XLMR*	*ApTCL*	*Sz*	*H*
	IFCL	*Ht*	*IO*	*AtTCL*	*CDLCL*
	VLC	*HpCL*	*NMC*	*LBC*	*APC*
	FT	*FO*	*GsM*	*S*	*AF*
	HLV	FTCL	*ADP*	*WLR*	*ANB*
	AS	IT	*NC*	*HF*	*AG1PTTCL*
	EOMO	AS	*DbM*	*E*	*AISCCL*
	FD	EAEM	*ALCL*	*AMC*	*DfM*
Pathway	*PTENS*	CCC	CCG2/M	**Wnt/BCS**	*NGB*
	PI3K	PPARS	PPARS	*CCC*	AP
	XMS	TyrM	ASM	GPCRS	
	GB		NOSCVS		
	SRS		GD		
	PPARS		p38MAPKS		
	IL10S		DGDGM		
	PD				
	EGFS				
	IGF1S				
Drug Trgts	Cg	Cg	None	ADAT	6RT
	NR	DS		RDT	PE1
		SBAV		ST	UV
		TPI			
		ZK			

Key: Phenotype is the gene whose expression values were most significantly associated with the SNP listed in the second row based on the analysis by Cheung et al. (2005); groups that were significantly overrepresented among the genes whose expression values were associated with each SNP are listed with p<.05: normal font; p<.01: italicized; p<.001: bold, and p<.0001: bold and italicized. Drug Trgts = Drug Targets.

Table 1. Continued.

Phenotype	CTSH	RPS26	IRF5	S100A13	CPNE1
SNP rs#	rs1369324	rs2271194	rs2280714	rs3757791	rs6060535
Functions	**TSf1BS**	**ADTL**	**QCIB**	**MD**	*DSCGN*
	ASF1RE	**ADC**	**QC**	**IN**	*PA*
	CkCL	**ADT**	**QAA**	**FPp**	*RgT*
	DCa	**CGIBCL**	*GOC*	**CMBL**	*AAT*
	DHN	*CLDNA*	*SCA*	*CMKCL*	*QM*
	DSC	*DS*	*ACC*	*ADL*	*DGT*
	EjM	*OWPO*	*AdO*	*MAPC*	*ACAM*
	MF	*RcDNA*	*CNKTL*	*STS*	*QMPC*
	Q12SHA	*CGIECL*	*CNKTL*	*QCP*	*FlT*
	DPG	*BH*	*DDEP*	*QLT*	*TBc*
Diseases	*AOMS*	**ICT**	**HDR**	**HCH**	**EB**
	EmH	**ICC**	*LO*	*AOMS*	*GABEB*
	GA	**PT**	*DCVT*	*IHPC*	*ApMb*
	HDH	*PcC*	*DN*	*KF*	*EBS*
	XPCGE	*VIP*	*AJ*	*LH*	*LoBC*
	DAN	*HFC*	*QPC*	*LN*	*MMGT*
	HR	*HEG*	*DC*	*CDSCC*	BM
	TCC	*BAC*	*DR*	*MSM*	CFCL
	IC	*IvTL*	*DCCL*	ADHD	CRBC
	HTg	*HDC*	*MTTCL*	CCaC	DO
Pathway	*IL4S*	**ButM**	*CaS*	OCPF	PTTB
	IL2S	*VLID*	PheM	SRS	ERSP
	NERP	*BAS*		PyrM	
	PDGFS	FAM		LD	
		RM		GPM	
		FAB2		OGB	
		LD		CCC	
		SDKB		PS	
		PM		PM	
		TM			
Drug Trgts	ATAC	Est	CTA	IDN	DAAT
	EIT	DAAT	CADV		Cg
	IDN	BFNT			FP
	NZ				IT
	PVF				INO
	RDT				OT
					YM

Key: See first part of the table.

Table 1. Continued.

Phenotype	TCEA1	IL16	SMARCB1	EIF3S8	CSTB
SNP rs#	rs6562160	rs6957902	rs7802273	rs8092794	rs880987
Functions	**ISRE**	***TCT***	***QCFEC***	***CPBM***	**AMt**
	APMP	**BP**	**APTP**	**SLE**	*QN*
	JDNAF	**SP**	*RCAMP*	**CMPBL**	*M*
	QTC	**FuL**	*CSC*	*CMp*	*CA*
	CCH	*TB*	*TRunx2*	*CaMo*	*DDSE*
	HRDNA	*BF*	*BOLA*	*MCm*	*ECRE*
	CmRNA	*FuP*	*LLs*	*CMTL*	*FTM*
	DRG	*IDNA*	*TCA*	*TPR*	*MSMC*
	AWAT	*TET*	*CR*	*CMc*	*NRGC*
	AAtf1BS	*BMt*	*LR*	*ICTL*	*SRM*
Diseases	**SBCCL**	*K*	*HTN*	**IHIV1**	*CVCCCL*
	TF	*CMMTC*	*PTTCL*	*CLCL*	*ARMCL*
	GS	*MMTC*	*HGC*	*CMLCL*	*ASGC*
	FIT	*ICCCL*	*ABL*	*E*	*AT*
	RgC	*EN*	*CHCL*	*APCCL*	*CGICCCL*
	AnT	ASMC	*FHC*	*ARBS*	*MDV*
	SCCL	ACCCL	*SSCCL*	*CAAM*	*Hc*
	GBC	CGICvC	*CVD*	*CAAM*	*HtM*
	PPCCL	DDMD	DDG	*CSNB1*	HSS
	ApEC	DA	FCNHL	*EM*	LN
Pathway	*CysM*	CysM	SB	GRS	*Wnt/BCS*
	PDGFS	VEGFS	CC	cAMPMS	*UB*
	EGFS	PTTB	BGGBN	CS	
	NOSCVS	CaS	BAM		
	AP	CCG1/S			
	FAB2				
	VEGFS				
	E/M				
	N/TrkS				
Drug Trgts	RF	DACA	Ac	A	Est
	NCF	MAD		AFD	DAAT
		NR		AVP	AVPL
		PTR		CL	BFNT
		PVF			KC
		SISB			

Key: See first part of the table.

Table 1. Continued.

Phenotype	CHI3L2	PPAT	PPAT	LRAP	HSD17B12
SNP rs#	rs755467	rs2139512	rs227940	rs2762	rs4755741
Functions	**MCC**	***Ct***	*DPG*	***TMA***	*QIP*
	APPC	***QC***	*BTL*	**IL**	*RMD*
	DHT	**CEkC**	*PHITL*	*QPBL*	*QBE*
	FP	**HEC**	*QC2*	*CFPBL*	*ATI*
	MC	**RSMC**	*BBrCL*	*ITL*	*AKC*
	AHNF4	**HCL**	*ON*	*GOC*	*SC*
	RCt	**Mca**	*QC*	*QPD*	*SBF*
	RCP	**SE**	*CBV*	*FcL*	*ATL*
	QOF	**RC**	*DvS*	*QGMPC*	*EESC*
	DSS	*REC*	*GPSC*	*RSMC*	*FEF*
Diseases	**PIN**	***IDS***	**HpCL**	***DBBB***	**RT**
	DE	**InfS**	*CHB*	**HpCL**	*TLCCL*
	DOP	**PH**	*HBCL*	**PMNM**	*TO*
	MCCL	**TAPC**	*NSCLC*	**HBCL**	*HiM*
	PSCD	**AICL**	*AVB*	*PR*	*C*
	PINM	**DNCL**	*AM*	*IF*	*G2/MP*
	PINM	*AECL*	*SADM*	*PdL*	*IICL*
	Hg	*Ha*	*HDC*	*SL*	*LB*
	ApO	*BA*	*N*	*AgTCL*	*MFCL*
	ApSc	*DNDM*	*CVCRCCL*	*AgBCCL*	*OB*
Pathway	*AtRNAB*	*GPM*	GPM	AP	CaS
	ERS	*BGGBN*	NKCS	PI3K	IL6S
	TGFBS	*PD*		PTENS	
	PurM	IRS		MM	
	GSTM	GluM		ASM	
		NgM			
		SRS			
		CBAS			
		NOSCVS			
		AP			
Drug Trgts	BP	AET	Est	EMD	FCTF
	MH	AVP	AS	EZ	FP
		DP	CTA	IGF1	XV
		DAA	EZ	TH	
		TAK	IGF1		

Key: See first part of the table.

Table 1. Continued.

Phenotype	PSPHL	HLA-DRB2	AA827892	CGI-96
SNP rs#	rs6593279	rs6928482	rs788350	rs9600337
Functions	**CDNA**	**DBMD**	*ECL*	**BBCL**
	MI	**GMC**	*PCL*	**CCPBCL**
	MA	**CVESC**	*DBCL*	**RP**
	ADPRAA	**RO**	*AB*	*FSR*
	CL	**StBMC**	*BCD28R*	*ARNA*
	HTb	*UPRC*	*DDLSC*	*ACCPCL*
	MP	*MN*	*FAA*	*CTm*
	PL	*SBL*	*LTL*	*Q12HA*
	SMXC	*CFCFUM*	*RLC*	*BLCL*
	DAF	*ELP*	*TCD28*	*ITP*
Diseases	**PC**	***HDC***	**PC**	**GPT**
	ACPBCL	**RUD**	**PBCCL**	**BCC**
	QSC	**HTgR**	*ETCL*	**FC**
	PCn	**RUDM**	*ASCCL*	*GLCCL*
	EGT	**P**	*CSCCL*	*GC*
	GP	**FO**	*DST*	*ATCL*
	LBBB	**F**	*ILCL*	*GMT*
	SCLL	**FL**	*HBMC*	*ALCCL*
	TC	**ApMC**	*Dac*	*AO*
	SM	**HDEC**	*DHCL*	*MLCCL*
Pathway	IRS	*Wnt/BCS*	NS	*FAB1*
	FAB1	ES	TCRS	*IS*
	GSTM	BAS	BCRS	*PI3K*
	IL4S		IM	SAPK/JNK
			GMCSFS	SHS
			SM	CCG1/S
			XMS	OCPF
				BCRS
				E/M
				FGFS
Drug Trgts	EPB	IMSR	CaT	LY
	IDN	PT	EDA	
	INO		FD	
	PVF		GC	

Key: See first part of the table.

Table 1. Continued; Control SNPs.

Phenotype	CXCL11	RIPK1	PLSCR1	PLSCR1
SNP rs#	rs10017431	rs10498658	rs2688692	rs2587021
Functions	*CCL*	**PBC**	**Ej**	*DH*
	CECL	**CTh2L**	*PCCL*	*RH*
	CNTCL	*AEC*	*AgEC*	*AN*
	NPC	*TOA*	*BATRE*	*DGC*
	BmRNA	*CAC*	*CEC*	*Ej*
	MLC	*EAC*	*EE*	*TTDNA*
	QIJ	*G0/SPT*	*PPF*	*RPA*
	PcP	*GUBC*	*SBV*	*G*
	AECC	*PP*	*AN*	*DD*
	CGJ	*CNKC*	*DGC*	*RGFAA*
Diseases	*DMC*	*RTCL*	***MLCL***	**ARVD**
	HpC	*ASCCCL*	**U**	**VTH**
	HAM	*IvLCL*	**AxR**	*El*
	IEC	*MHCL*	*AxM*	*IDDMM*
	GnM	*RCCCL*	*HS*	*AxR*
	EP	HB	*IK*	AtR
	NM	HCb	*FHM*	CMG
	PcR	JRA	CM	CDO
	SADM	MGC	CMCC	Hm
	Hp	SCTCL	AlR	VT
Pathway	IFS	*AtRNAB*	CS	GABARS
		PTTB	VEGFS	AAM
		SCLB	cAMPMS	cAMPMS
		FAB2		ES
Drug Trgts	Est	FP	Ac	RS
	AAH	IG1B	CB	SR
	BR	SM	EZR	TG
	IM		NN	
	IMSR		RZ	
	OM		TH	
	RGPT			

Key: See first part of the table; Note that control SNPs were not significantly associated with their local genes in t-tests.

Table 2. Probability of True Positive and Expected Discovery Rate Analysis of Gene Expression Associations with Each SNP

SNP RS#	# p<0.05	PTP	EDR	# p<0.01	PTP	EDR
rs10490570	1117 (13.1)	0.80	0.80	568 (6.7)	0.94	0.70
rs10509846	1081 (12.7)	0.84	0.65	337 (4.0)	0.95	0.44
rs1060043	834 (9.8)	0.57	0.32	197 (2.3)	0.65	0.09
rs10807387	1126 (13.2)	0.71	0.43	391 (4.6)	0.85	0.19
rs11822822	724 (8.5)	0.28	0.14	150 (1.8)	0.25	0.14
rs1369324	560 (6.6)	0.45	0.29	146 (1.7)	0.68	0.16
rs2271194	400 (4.7)	0.02	0.01	94 (1.1)	0.01	0.00
rs2280714	535 (6.3)	0.59	0.12	124 (1.5)	0.71	0.04
rs3757791	553 (6.5)	0.41	0.63	187 (2.2)	0.73	0.48
rs6060535	519 (6.1)	0.25	0.35	134 (1.6)	0.44	0.16
rs6562160	753 (8.8)	0.46	0.22	188 (2.2)	0.50	0.05
rs6957902	1051 (12.3)	0.51	0.19	281 (3.3)	0.52	0.04
rs7802273	1371 (16.1)	0.77	0.38	509 (6.0)	0.87	0.14
rs8092794	746 (8.8)	0.63	0.51	247 (2.9)	0.83	0.30
rs880987	662 (7.8)	0.35	0.16	176 (2.1)	0.36	0.04
rs755467	466 (5.5)	0.07	0.10	110 (1.3)	0.04	0.01
rs2139512	799 (9.4)	0.38	0.15	208 (2.4)	0.34	0.02
rs227940	373 (4.4)	0.04	0.01	67 (0.8)	0.01	0.00
rs2762	319 (3.7)	0.18	0.01	56 (0.7)	0.09	0.00
rs4755741	657 (7.7)	0.57	0.69	245 (2.9)	0.84	0.56
rs6593279	466 (5.5)	0.17	0.06	101 (1.2)	0.16	0.01
rs6928482	628 (7.4)	0.51	0.32	177 (2.1)	0.69	0.14
rs788350	541 (6.3)	0.32	0.43	145 (1.7)	0.56	0.23
rs9600337	1358 (15.9)	0.87	0.76	597 (7.0)	0.96	0.60
rs10017431	379 (4.4)	0.00	0.00	82 (1.0)	0.00	0.00
rs10498658	413 (4.8)	0.00	0.00	80 (0.9)	0.00	0.00
rs2688692	322 (3.8)	1.00	0.02	51 (0.6)	1.00	0.00
rs2587021	253 (3.0)	1.00	0.01	34 (0.4)	1.00	0.00

Key: # p<x is the number of genes whose association strength with the SNP designated in the left most column produced a p-value less than x (numbers in parentheses designate the percentage of the total number of genes with p<x); PTP is the probability of a true positive result; EDR is the expected discovery rate.

Table 3. Multivariate Distance Matrix Regression Analysis of SNPs Influencing the Serotonin Receptor and the Wnt/Beta-Catenin Signaling Pathways.

		Marginal results		Conditional results		
Pathway	**SNP**	**p-value**	**%variation**	**p-value**	**%variation**	**cumulative**
Serotonin receptor	rs2139512	0.0109	5.0	0.0109	5.0	5.0
signaling	rs10490570	0.0294	4.1	0.0311	3.8	8.8
	rs3757791	0.1934	2.5	0.5421	1.4	10.2
Wnt/Beta-catenin	rs10807387	0.0008	5.0	0.0008	5.0	5.0
signaling	rs10509846	0.1350	2.5	0.1118	2.5	7.4
	rs880987	0.4037	1.8	0.2318	2.1	9.5
	rs6928482	0.2185	2.2	0.4961	1.6	11.1

Key: Marginal results: an analysis of each SNP tested independently; Conditional results: forward stepwise regression results. % variation: percentage of variation in the similarity matrix explained by the SNP; cumulative: cumulative percentage of variation explained the SNP(s).

Table 4. Multivariate Distance Matrix Regression Analysis With the Subset of Genes Associated With Each SNP Influencing the Wnt/Beta-Catenin Signaling Pathway.

		Marginal results		Conditional results		
Subset of genes (SNP)	**SNP**	**p-value**	**%variation**	**p-value**	**%variation**	**cumulative**
rs10509846	rs10509846	0.0369	4.59	0.0321	4.54	9.26
	rs10807387	0.0356	4.71	0.0356	4.71	4.71
rs10807387	rs10807387	0.0001	6.9	0.0001	6.9	6.9
rs880987	rs880987	0.0108	4.94	0.0108	4.94	4.94
rs6928482	rs6928482	0.0084	5.29	0.0084	5.29	5.29

Key: Note: All four SNPs were used in analyses, but only significant SNPs are shown. Marginal results: an analysis of each SNP tested independently; Conditional results: forward stepwise regression results. % variation: percentage of variation in the similarity matrix explained by the SNP; cumulative: cumulative percentage of variation explained the SNP(s).

ACKNOWLEDGEMENTS

The text of Chapter 4 has been submitted for publication as:

> Jennifer Wessel, Matt Zapala and Nicholas J. Schork. Accommodating Pathway Information in Expression Quantitative Trait Locus ("eQTL") Analysis

The dissertation author was the primary researcher and/or author and the co-authors listed in this publication directed and supervised the research which forms the basis for this chapter.

CHAPTER 5

Conclusions and Future Directions

The aims of the three studies described in this dissertation were to develop an analysis method that is appropriate for handling the massive amounts of genetic information that genetic epidemiologists will have at their disposal for large-scale association studies. The method was inspired by an acceptance of many biological realities associated with the human genome, such as its diploid nature, the fact that genes do not work in isolation to influence a phenotype, and that the exploitation of linkage disequilibrium (LD) information is, although necessary with today's technologies, not likely to be necessary in the future when complete DNA sequence data may be available on individuals in a study. The proposed method has been extended and applied to scenarios of great interest in contemporary genetic epidemiology research; e.g., whole genome associations and a pathway-centric approach to genetic association analyses. Each chapter in the thesis describes the merits and limitations of the studies taking advantage of the approach, as described in chapters 2–4. In the following I provide an overview of the main findings of my research and consider its limitations and areas for future research.

Main Findings

The development and application of the multivariate distance matrix regression (MDMR) method to public data demonstrated the utility of the approach to genetic associations involving multilocus genotype data (Chapter 2). The method was shown to be able to accommodate weightings factors that consider 'functional' information about genetic loci. Such weighting schemes were shown to improve the association analysis results. The method was applied to the analysis of gene expression data obtained on accessible tissues from a sample of humans. Gene expression data of this sort are being studied more and more often in clinical and field epidemiologic studies of disease in order to identify biomarkers associated with that disease. Hence, my association studies using the MDMR method did not only address methodological and practical concerns with association studies, but also with the analysis of state-of-the-field biomarker assay results.

The MDMR method was also shown to have utility in whole genome association (WGA) studies (Chapter 3). The application of the MDMR method identified a number of loci potentially influencing CHI3L2 gene expression levels. In order to determine how these loci could be working together to influence the phenotype, their interactions were explored and revealed that a number of these genes appear to work together at the molecular level. This led me to consider how SNPs in multiple genes, possibly implicated in a single pathway, could collectively influence a complex phenotype (Chapter 4). I was able to show that individual SNPs may be associated with the expression levels of a large number of genes, and thus many naturally occurring genetic variations may actually influence the expression of many genes in known biochemical networks. Thus I was able to show not only that multiple SNPs can influence a common genetic or biochemical pathway, but multiple pathways may be influenced by a single SNP, thus demonstrating the very great potential for genetic heterogeneity in the mediation of complex human traits and diseases.

Although I focused much of my research effort on the analysis of multiple genetic factors in different contexts, there are a number of additional epidemiologic issues that demand consideration in genetic epidemiologic association studies that often, unfortunately, are ignored. For example, the influence of non-genetic factors, such as environmental or epigenetic factors or their combination, on the relationship between a particular genetic variation and a trait should be assessed and controlled for in relevant analyses and/or study design. In addition, sampling biases can and often do plague genetic association studies, which are often based on convenience samples or sampling strategies that unknowingly (or knowingly) sample cases and controls from different source populations, which can lead to substantial false positive and false negative rates in genetic association studies [1-19]. Misclassification bias due to genotyping error from allele calls in current high-throughput methods could lead to misclassifying individual genotypes. The presence of misclassification bias will more likely mask or reduce associations, than to cause them. In this context the International HapMap Project investigators went to extensive efforts to increase the reliability of genotype calls to avoid bias in the study. Using large numbers of polymorphisms in a genetic epidemiologic study has the added benefit of determining the relatedness of individuals which in many standard epidemiologic study designs (prospective/cohort, case-control or cross-sectional) are assumed to be unrelated.

The accurate measurement of phenotypes, especially molecular phenotypes designed to act as biomarkers, such as gene expression levels and the reliability of the type of tissue can impact effect estimates or the correct interpretation of the physiologic process. Different tissue types will express different genes due to their specialization of the according cell types in the tissue [20, 21]. Therefore, only certain pathways can be found in certain tissues. Thus it might be that the pathways I identified in these studies will be different from pathways I would identify, if we had used a different tissue source to measure the expression levels of genes. Even more, the expression levels of genes can depend on external or internal circumstances, disease or exposure respectively (Tables 2 and 3, Chapter 1). A few studies have identified specific genes that are induced by the transformation process used in this study 22-24 [1-3]

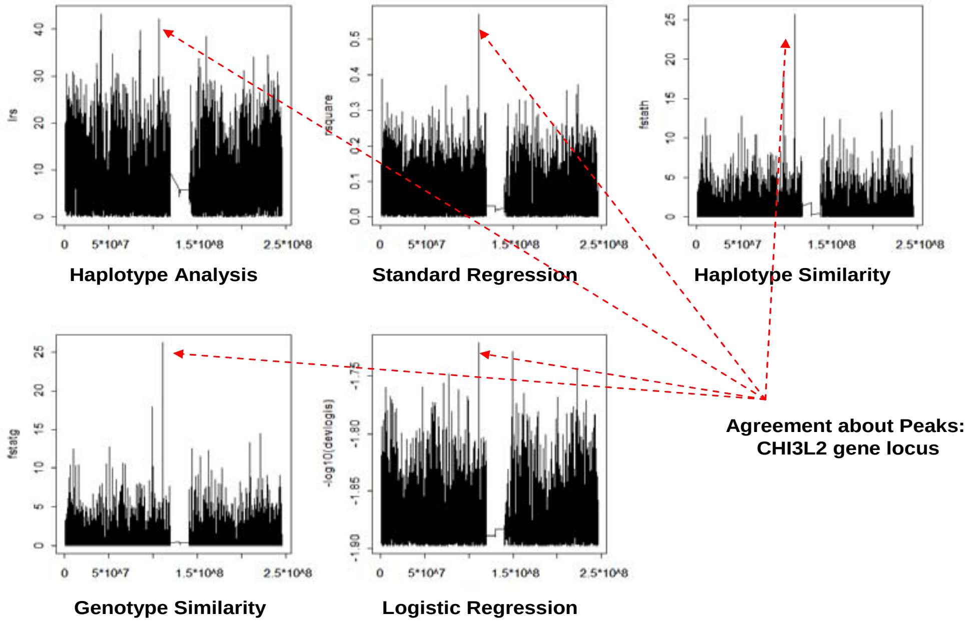
Many study populations are of European descent, which can limit external validity, but can reduce possible confounding by population stratification. The CEPH individuals represent individuals of European descent and are Mormons from Utah, sampled for their large, multi-generational pedigrees http://www.cephb.fr/. In this study, to assume individuals were unrelated and independent for statistical tests, only the parents from the trios chosen by the HapMap were used. Gender was represented equally (28 (49.1%) males and 29 (50.9%) females)). Age was available on 39 (68.4%) of the individuals and the mean age for these individuals is 73.9±9.8 years. For those missing age data it is difficult to estimate their possible ages since both parents are missing, but since parents of adult children were chosen their ages are most likely similar to those with available data. Unfortunately no other descriptive variables are available on this public data, most likely to assure anonymity.

The method is applicable to any genetic association study but the results may differ by study population, for example, because of known differences in allele frequency by ethnicity or geographic location, or by the different tissues used. The MDMR method is very general and can incorporate covariates, such as age, often a predictor of outcomes.

Many of these issues have been considered as a reason for inconsistent results in genetic associations [7, 25-35], and are starting to be addressed by the research community via efforts to, e.g., collect prospective data (for, e.g., gene x environment studies), establish initiatives such as the “Network of Investigator Networks” started by investigators at the CDC to share tools and methods of genetic epidemiology [36], criteria for journal reviews and manuscripts that incorporate appropriate guidelines [34].

Future Research in Genetic Epidemiology

Although I have shown that the MDMR method has great utility in modern genetic association analysis – especially those considering modern biomarkers – there are further research questions that should be considered in order to determine the utility of the method in other study design and biological effect scenarios. My colleagues and I have been comparing the MDMR method to other association analysis approaches, such as haplotyping and regression methods (e.g., logistic regression and linear regression). See Figure 1 below which compares the results of two versions of the proposed MDMR procedure with haplotype and regression procedures on the chromosome 1 data used in this dissertation. The results of the studies are being written up for publication (Multiple Regression Analysis in Candidate Gene and Whole Genome Association Studies. N Malo, J Wessel, N Schork, in preparation). We plan to apply the MDMR method to larger datasets, which only recently became available, that consists of ~50,000 gene expression values and the recent release of HapMap data (October 2006). This analysis will provide a more complete picture of the genes involved in gene expression characterized biochemical pathways. In addition, we plan on pursuing extensive simulations studies to completely evaluate the utility of the MDMR and related association analysis methodologies.



Haplotype Analysis

Standard Regression

Haplotype Similarity

Genotype Similarity

Logistic Regression

Agreement about Peaks: CHI3L2 gene locus

Figure 1. Comparison of Analysis Methods by Test-Statistic and Genomic Location on Chromosome 1

Genetic epidemiologists have entered an exciting era for understanding the genetic basis of common diseases. The notion that relevant research questions are deemed as so important that students in the epidemiological sciences should consider them in their doctoral work is becoming a reality. For example, Nicholas Schork, my doctoral thesis advisor, wrote his own doctoral thesis dissertation in an epidemiology program that considered theoretical and applied aspects of genetic studies that used microsatellites in pedigree-based linkage analysis contexts [38]. Since then, many students in epidemiological sciences have pursued genetic analysis methodology and applications-oriented thesis projects.

What is more, large-scale genetic epidemiology studies focusing on disease are becoming more and more popular. For example, in 2000, Jeanette McCarthy, a member of my thesis committee, embarked on the ‘GeneQuest’ of cardiovascular disease – a study designed to genotype thousands of individuals on >200 SNPs in cardiovascular disease candidate genes, and was considered one of the largest studies of its kind at that time [39]. Now in 2006, researchers have already performed genome-wide association analyses with as many as ~1 million loci [40-43]. As mentioned in the introduction, efforts at the national level are being undertaken to ramp such studies to a larger scale, with larger sample sizes (1000’s of individuals), more genetic loci, and the use of molecular phenotypes. These studies will begin to ask questions beyond whether a genetic factor is simply associated with a disease, but precisely what mechanisms might be involved and what clinical and public health consequences the findings such studies might have. Complementary to these studies are efforts to basically understand how humans have evolved with and without disease at the genomic levels (see, e.g., the large study being undertaken by the National Geographic Society and IBM: https://www3.nationalgeographic.com/genographic/).

Whatever the mission of these genetic studies, the need for a coherent, flexible, biologically-intuitive set of analysis approaches will be needed. My research was an overt attempt to devise such tools, implement them, and apply them to data whose results could be interpreted for their biological meaningfulness.

REFERENCES

1. Bacanu SA, Devlin B, Roeder K (2002) Association studies for quantitative traits in structured populations. Genet Epidemiol 22: 78-93

2. Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66: 1933-44

3. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004

4. Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20: 4-16

5. Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. Am J Hum Genet 71: 304-11

6. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36: 388-93

7. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598-604

8. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512-7

9. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68: 466-77

10. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of Confounding of Genetic Associations in Stratified Populations. Am J Hum Genet 72

11. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74: 965-78

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

13. Thomas DC, Witte JS (2002) Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 11: 505-12

14. Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60: 227-37

15. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220-8

16. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170-81

17. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63: 1839-51

18. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274-86

19. Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60: 957-64

20. Du X, Tang Y, Xu H, Lit L, Walker W, Ashwood P, Gregg JP, Sharp FR (2006) Genomic profiles for human peripheral blood T cells, B cells, natural killer cells, monocytes, and polymorphonuclear cells: comparisons to ischemic stroke, migraine, and Tourette syndrome. Genomics 87: 693-703

21. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N, Khan J (2005) Database of mRNA gene expression profiles of multiple human organs. Genome Res 15: 443-50

22. Zhao B, Maruo S, Cooper A, M RC, Johannsen E, Kieff E, Cahir-McFarland E (2006) RNAs induced by Epstein-Barr virus nuclear antigen 2 in lymphoblastoid cell lines. Proc Natl Acad Sci U S A 103: 1900-5

23. Maier S, Staffler G, Hartmann A, Hock J, Henning K, Grabusic K, Mailhammer R, Hoffmann R, Wilmanns M, Lang R, Mages J, Kempkes B (2006) Cellular target genes of Epstein-Barr virus nuclear antigen 2. J Virol 80: 9761-71

24. Li Y, Mahajan NP, Webster-Cyriaque J, Bhende P, Hong GK, Earp HS, Kenney S (2004) The C-mer gene is induced by Epstein-Barr virus immediate-early protein BRLF1. J Virol 78: 11778-85

25. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4: 45-61

26. Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87: 4438-41

27. (1999) Freely associating. Nat Genet 22: 1-2

28. Bird TD, Jarvik GP, Wood NW (2001) Genetic association studies: genes in search of diseases. Neurology 57: 1153-4

29. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2: 91-9

30. Altshuler D, Kruglyak L, Lander E (1998) Genetic polymorphisms and disease. N Engl J Med 338: 1626

31. Tabor HK, Risch NJ, Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3: 391-7

32. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306-9

33. Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, Cucca F, Guja C, Ionescu-Tirgoviste C, Stevens H, Carr P, Nutland S, McKinney P, Shield JP, Wang W, Cordell HJ, Walker N, Todd JA, Concannon P (2002) Parameters for reliable results in genetic association studies in common disease. Nat Genet 30: 149-50

34. Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG (2003) Genetic associations in large versus small studies: an empirical assessment. Lancet 361: 567-71

35. Ordovas JM (2003) Cardiovascular disease genetics: a long and winding road. Curr Opin Lipidol 14: 47-54

36. Ioannidis JP, Gwinn M, Little J, Higgins JP, Bernstein JL, Boffetta P, Bondy M, et al. (2006) A road map for efficient and reliable human genome epidemiology. Nat Genet 38: 3-5

37. (2005) Framework for a fully powered risk engine. Nat Genet 37: 1153

38. Schork NJ. Advances in the Genetic-Epidemiologic Analysis of Complex Phenotypes. *Department of Epidemiology*. Ann Arbor: University of Michigan, 1994; 142.

39. Topol EJ, McCarthy J, Gabriel S, Moliterno DJ, Rogers WJ, Newby LK, Freedman M, Metivier J, Cannata R, O'Donnell CJ, Kottke-Marchant K, Murugesan G, Plow EF, Stenina O, Daley GQ (2001) Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. Circulation 104: 2641-4

40. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78

41. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38: 617-9

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

43. Wessel J, Libiger O, Schork NJ (2006) Whole Genome Association Studies Using Window-Based Multivariate Distance Matrix Regression Analysis. Genetic Epidemiology Submitted.