UNIVERSITY OF CALIFORNIA, SAN DIEGO

Neurogenomics in the mouse model: multivariate statistical methods and analyses

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Biomedical Sciences

by

Matthew Alan Zapala

Committee in charge:

Professor Nicholas J. Schork, Chair
Professor Philip E. Bourne
Professor Daniel T. O'Connor
Professor Bernhard Ø. Palsson
Professor Anthony Wynshaw-Boris

2007

The Dissertation of Matthew Alan Zapala is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2007

DEDICATION

This work is dedicated to my teachers, particularly Dr. Schork who has been instrumental in the completion of this body of work, to my parents and lastly to my wife and son whose support has been invaluable to me.

EPIGRAPH

We are drowning in information and starving for knowledge.

Rutherford D. Roger

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

LIST OF TABLES

Chapter 4

Chapter 5

Chapter 6

Chapter 7

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Nicholas J. Schork whose mentoring, guidance and support have provided me with the research tools to succeed. I want to gratefully acknowledge the committee members Dr. Anthony Wynshaw-Boris, Dr. Daniel O'Connor, Dr. Philip Bourne and Dr. Bernhard Palsson for their guidance in these projects.

I would also like to acknowledge the entire Schork Laboratory for their contributions to the body of work herein. Particularly, Ondrej Libiger who assisted me in programming two of the statistical applications and Charles Abney who assisted in web development for these same programs.

Also, I would like to acknowledge Carrolee Barlow, David Lockhart, Iiris Hovatta, Jennifer Greenhall, Chun (Jimmie) Ye and Eleazar Eskin whose collaborations have played a significant role in this dissertation.

The text of Chapters 2, 3, 4 and 5 are a formatted reprint of material which appeared in journals *Proceedings of the National Academy of Science, USA*, *Genome Biology* and *Genome Research.* Full citations follow:

Zapala MA*, Hovatta I*, Ellison JA*, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Reilly JF, Young WG, Bloom FE, Lockhart DJ, Barlow C. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*. 2005. 102:10357-62.

Hovatta I*, **Zapala MA***, Broide RS, Schadt EE, Libiger O, Schork NJ, Lockhart DJ, Barlow C. DNA variation and brain-region specific expression profiles show different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biology*. 2007. 8:R25.

Greenhall JA*, **Zapala MA***, Caceres M, Libiger O, Barlow C, Schork NJ, Lockhart DJ. Detecting genetic variation in microarray expression data. *Genome Research*. 2007. In Press.

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A*. 2006. 103:19430-5.

The dissertation author was the primary author. The co-authors listed in the chapters either assisted, directed or supervised the research that forms the basis for this dissertation.

VITA

1998	Bachelor of Science, Biological Sciences Stanford University Departmental Honors
1997	E.O. Research Fellow, Lawrence Berkeley National Laboratory Department of Genetic Toxicology Laboratory of Dr. Susan Anderson
1998-2000	Environmental Scientist Kinnetic/Toxscan Laboratories Laboratory of Dr. Patrick Kinney
2000-2002	Research Assistant II, Salk Institute for Biological Studies Department of Genetics Laboratory of Dr. Carrolee Barlow
2007	Doctor of Philosophy, Biomedical Sciences University of California, San Diego
2009	Doctor of Medicine, School of Medicine University of California, San Diego

TEACHING

2006	UCSD School of Medicine, Department of Surgery, Human Anatomy Course, Teaching Assistant

COMMITTEES/ORGANIZATIONS

2002-2004	Introduction to Clinical Medicine, Student Course Representative
2003-2006	UCSD School of Medicine Alumni Board, Student Representative
2004-2005	UCSD MSTP (M.D./Ph.D.) Admissions Committee Member
2004-2005	UCSD Committee on Academic Information Technology, Graduate Student Representative

PUBLICATIONS

Greenhall JA*, **Zapala MA***, Caceres M, Libiger O, Barlow C, Schork NJ, Lockhart DJ. Detecting genetic variation in microarray expression data. *Genome Research*. 2007. In Press.

Wessel J, **Zapala MA**, Schork NJ. Accommodating pathway information in expression quantitative trait locus (eQTL) analysis. *Genomics*. 2007. In Press.

Hovatta I*, **Zapala MA***, Broide RS, Schadt EE, Libiger O, Schork NJ, Lockhart DJ, Barlow C. DNA variation and brain-region specific expression profiles show different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biology*. 2007. 8:R25. **Highly Accessed**.

Zapala MA, Barlow C, Hovatta I. Molecular Anatomy of the Mammalian Brain. In Bloom F, Gage F, Squire L (Eds.). *New Encyclopedia of Neuroscience*. 2007. New York, Elsevier. In Press.

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A*. 2006. 103:19430-5.

Zapala MA*, Hovatta I*, Ellison JA*, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Reilly JF, Young WG, Bloom FE, Lockhart DJ, Barlow C. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*. 2005. 102:10357-62.

Caceres M, Lachuer J, **Zapala MA**, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A*. 2003. 100:13030-5.

Zapala MA, Lockhart DJ, Pankratz DG, Garcia AJ, Barlow C, Lockhart DJ. Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biology*. 2002. 3:1-9.

MANUSCRIPTS IN SUBMISSION

Ye C*, **Zapala MA***, Min Kang H, Wessel J, Eskin E, Schork NJ. High-Density QTL Mapping to Identify Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways. *BMC Genomics*. 2007.

Zapala MA, Schork NJ. Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis. *Communication in Statistics: Simulation and Computation*. 2007.

PUBLISHED ABSTRACTS

Wachsman W, **Zapala MA**, Hata T, Walls L, Wong R, Schork NJ, Chang S. Differentiation of melanoma from dysplastic nevi in suspicious pigmented skin lesions by non-invasive tape stripping. *Society for Investigative Dermatology*, 2007. Best Clinical Application Plenary Poster, awarded by Nature Publishing.

Luo W, Ng S, Schork NJ, Hermann TW, Zhang JK, **Zapala MA**, Meglasson MD, Negro-Vilar A. A genomewide association study to identify genomic biomarkers of Bexarotene induced hypertriglyceridemia and longer survival in NSCLC. *American Society of Clinical Oncology*, 2007.

Ye C*, **Zapala MA***, Kang HM, Wessel J, Eskin E, Schork NJ. High-Density QTL Mapping to Identify Phenotypes and Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways. *Second RECOMB Satellite Workshop of Systems Biology*, 2006.

Zapala MA, Wessel J, Schork NJ. High-density QTL mapping to identify loci influencing gene expression patterns in entire biochemical pathways. *American Society of Human Genetics*, 2006.

Zapala MA, Hsaio G, Chang KN, Schork NJ. EvolGenomic: a tool for translational genomics. *American Society of Human Genetics*, 2005.

Ellison JA, Wodicka L, Del Rio JA, Hovatta IM, Tyan W, Tennant R, Winrow C, Broide R, Reilly J, **Zapala MA**, Young W, Bloom F, Lockhart DJ, Barlow C. Molecular technology meets neurobiology. *Society for Neuroscience*, 2004.

Wright G, **Zapala MA**, Davis R, Bagul A, Travis A, Stone L, Shenoy S, Rote M, Glick R, Allingham D, Shenoy S, Del Rio J, Lockhart DJ and Barlow C. TeraGenomics: Scalable Computing for Genomic Neuroscience. *Society for Neuroscience*, 2003.

PRESENTATIONS

Zapala MA. Neurogenomics in the mouse model: statistical analyses and translational considerations for neurological phenotypes. National Public Health Institute, Helsinki, Finland, 2006.

Zapala MA. Bioinformatic applications for translational genomics. San Diego Supercomputer Microarray Seminar, 2005.

Zapala, M.A. Microarray data analysis on two platforms, cDNAs and Affymetrix: Revealing the black box. Day of Arrays Symposium, The Salk Institute for Biological Studies. 2001.

ABSTRACT OF THE DISSERTATION

Neurogenomics in the mouse model: multivariate statistical methods and analyses

by

Matthew Alan Zapala

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2007

Professor Nicholas J. Schork, Chair

The use of high-throughput genomic technologies has led to significant advances in the study of the molecular anatomy of the mammalian brain and the creation of the field known as neurogenomics. Microarray gene expression data has been utilized to identify genes associated with specific brain functions, behaviors and disease-related phenotypes. These gene expression datasets have shed light on the molecular organization of both the developing and adult mammalian brain, specifically in the mouse model. To investigate the molecular organization of the adult mammalian brain, a gene expression-based brain map was built. Gene expression patterns for 24 neural tissues covering the mouse central nervous system were measured and it was found, surprisingly, that the adult brain bears a transcriptional "imprint" consistent with both embryological origins and classic evolutionary relationships. Beyond simply analyzing gene expression patterns within the brain, it is now possible to analyze genomic sequence data, such as single nucleotide polymorphisms (SNPs), in parallel with large scale gene expression data in what has

been called a genetical genomics approach to determine transcriptional regulatory networks. To further analyze the molecular organization of the mouse brain, we analyzed gene expression profiles of five brain regions from six inbred mouse strains and integrated these findings with SNP data available for the individual strains. We found that many transcriptional regulatory networks are highly specific to particular brain regions. The ability to query the rich, complementary data sources of gene expression and SNPs together offers tremendous inroads to start to unravel the genetic determinates of complex polygenic diseases and phenotypes. However, appropriate data analysis strategies must be developed that can accommodate the complexity and high-dimensional aspects of these disparate data sources. In order to address some of these analysis issues, we developed an algorithm to identify sequence variation in gene expression data which can artificially affect expression signals and lead to false positive results. We also expanded a new statistical technique termed multivariate distance matrix regression that tests the association of multivariate profiles arising from high-dimensional data sets common in neurogenomics. The body of work presented herein attempts to assimilate the distinct fields of neuroanatomy, genomics, bioinformatics, statistical genetics and biostatistics to create novel analysis tools and develop new insights into biological processes related to neurogenomics.

CHAPTER 1

Introduction and Specific Aims

Introduction

The very heterogeneous nature of the structures and functions associated with the brain not only complicate studies of its origins both evolutionarily and developmentally, but also studies of brain-related phenotypes, such as neuropsychiatric diseases and complex behaviors. Thus, given the large number of genes encoded in the genomes of higher mammals, and the complexity of phenotypes associated with behavior, learning and memory, how is it possible to identify genes that regulate these phenotypes? How can we use genetic and genomic information to discover targets for therapeutic intervention of brain-related diseases? As genomic technologies, such as DNA microarrays and high-throughput *in situ* hybridizations, have become more commonplace, they have led to significant advances in the study of the molecular anatomy of the mouse, human and primate brain and, in fact, have paved the way for a new research field known as "neurogenomics" (Zapala et al. 2005; Lein et al. 2007). Neurogenomic researchers have utilized massive amounts of available genomic data to identify genes associated with specific brain functions, behaviors and disease-related phenotypes (Carter et al. 2001; Hovatta et al. 2005; Nadler et al. 2006). Emerging technologies, experimental designs, and data analysis methods have improved so that it is now possible to reproducibly measure gene expression levels in different regions, nuclei, and even single cells of the brain, and to reliably detect very subtle expression differences across individuals with different brain-related phenotypes (Lombardino et al. 2006).

An important goal in neuroscience is the development of a combined high-resolution map of the molecular, genetic and physical anatomy of the brain. This goal, although challenging, is now technically feasible (Zapala et al. 2007). The purpose of obtaining such a map and developing the tools to navigate it is to provide a more complete context for neurobiological studies in much the same way that the genome sequence provides a genetic context for the study of biology. But like sequence information, such a map is more than just a reference text and a trail guide; it is also a discovery tool that will help neuroscientists formulate and test hypotheses about the brain, speed the identification of genes important for brain function, assist in understanding the molecular and genetic building blocks of brain processes such as learning and memory, and facilitate the identification of the cause of certain CNS diseases.

Despite a great deal of initial skepticism, it is now clear that genomic approaches, such as parallel expression profiling with DNA microarrays, can be used to help elucidate the workings of the brain at the molecular level (Barlow and Lockhart 2002). Some of the skepticism concerning the application of genomic techniques to neuroscience was not wholly unwarranted, however, as the study of the brain brings a variety of somewhat unique and formidable technical challenges. Studies of the nervous system are complicated by the following: 1. the brain is a very heterogeneous organ with many different components; 2. the cells of greatest interest might comprise only a small fraction of the specific brain tissue of interest; 3. high-quality biomaterial can be hard to obtain; 4. the appropriate anatomical divisions

between regions are not always clear; 5. repeated dissections may be inconsistent; 6. some anatomical and functional regions are extremely small; 7. cell bodies are often located in a different part of the brain than the processes they may be associated with (axons and dendrites); 8. biologically important expression changes may be very subtle; and 9. gene expression profiles are dynamic and subject to change in response to even minimal perturbations. Nevertheless, there have been great strides made in performing genome-wide gene expression studies in the brain, as is detailed in this work.

Genetic and Genomic Analysis Approaches to the Study of the Brain

The completion of the Human Genome and Human HapMap Project has motivated researchers to consider the identification of the fundamental mutant or variant genes associated with many common human diseases (International Human Genome Sequencing Consortium 2001; International HapMap Consortium 2005). However, the question of how to approach identifying the specific genes and DNA sequence variations mediating disease susceptibility with the information provided by these initiatives is difficult due to the complexity and sheer amount of data. Moreover, the identification, collation, and structural annotation of genes and sequence variations from large-scale sequencing and genotyping efforts do not by themselves provide sufficient information about the functions or physiologic effects of genes. Thus, linking particular genetic variations to specific human diseases is far from easy. As a

result, many strategies have been devised to consider the functions of genes and the impact that variation in those genes might have on those functions.

Two common genomic approaches to identify functional genetic variants have been microarray gene expression analysis and quantitative trait loci (QTL) analysis. Gene expression analysis has been useful to identify differentially expressed genes between disease (i.e., case) and control groups (Alon et al. 1999) whose differences in expression may reflect true inherent biological differences between the two groups. However, numerous genes are usually identified as differentially expressed between the case and control groups. Isolating the exact causative genes responsible for the phenotypic differences between the two groups can be daunting.

Forward genetic approaches, such as QTL mapping and analysis, have been used to locate genomic regions that co-segregate with a phenotype of interest (Botstein et al. 1980). The distinction between forward and reverse genetic approaches lies in the starting point: forward genetic approaches seek to identify the genetic basis of a trait by focusing on genetic explanations for individuals exhibiting phenotypic variations, while reverse genetic approaches seek to characterize and categorize phenotypes that may have resulted from specific genetics changes (i.e., specific SNPs, insertions, deletions, repeat polymorphisms, etc.). Forward genetic studies have been used with some success to identify genes associated with overtly Mendelian and monogenic traits, such as Cystic Fibrosis, Huntington's disease and polycystic kidney disease (Gusella et al. 1983; Reeders et al. 1985; Riordan et al. 1989). However, forward genetics has been far less successful in identifying variants associated with

common complex traits and behaviors involving multiple genes and or environmental factors, such as high blood pressure, type II diabetes and major depressive disorder (An et al. 2005). One reason is that very large sample sizes are needed if the frequency of the disease or size of the genetic component is small (Schork 2002). Moreover, in human populations, clinical diagnoses can misclassify individuals and contaminate experimental groups (Schadt 2005). In addition, these studies often identify large genomic regions containing several hundred to several thousand genes and fine mapping to identify specific genes can be lengthy and arduous without guaranteed success. The study of model organisms, such as mice, has been considered as a viable alternative to human studies since experimental parameters, such as environment, breeding scheme, and phenotyping, can be controlled (Macauley and Ladiges 2005).

The Mouse as a Model Organism in Neurogenomics

Laboratory mice were developed approximately 100 years ago and have been the most widely used model organism for obtaining biological leads on human disease pathologies in ways not possible with traditional human clinical protocols (Wade and Daly 2005). As mammals, mice are physiologically and genetically more similar to humans than are traditional model organisms, such as flies, worms and yeast. The evolutionary distance between mice and humans is less than one-tenth of that of the next nearest model organism, *Drosophila melanogaster* (Boffelli et al. 2004). From a genetics viewpoint, mice are ideal mammals as it is straightforward to feed and house them, they can adapt to human manipulation and most importantly, they are easy to

breed with relatively short generation times. The ability to direct matings means that it is possible to create inbred strains, i.e., lines of mice that are genetically identical. In addition, powerful techniques have been developed to allow the manipulation of the mouse genome in multiple ways, such as creating gene knockouts, transgenic and consomic mice (Frankel 1995; Geyer et al. 2002; Accili 2004; Austin et al. 2004; Singer et al. 2004).

The ability to utilize mice as surrogates for human pathology is particularly critical in neuropsychiatric diseases as brain tissue from human patients is difficult to obtain. Also, misdiagnoses of psychiatric patients are potentially more common than in other human diseases, as patient findings are often purely clinical without quantitative measurements (Hyman 2002). Most importantly, from a neuropsychiatric viewpoint, mice have a central nervous system (CNS) and brain which can be genetically manipulated. However, the brains and behavioral patterns of mice and humans have evolved substantially since their divergence from a common ancestor approximately 75 million years ago (Zhang et al. 2004). The adult mouse brain has a mass that is less than 1/2000 of the human brain (Tecott 2003). The expansion in particular of the human cerebral cortex, which has a surface area a thousand times greater than that of the mouse, has given rise to the complexity of cognitive ability, emotional behaviors, and social interactions (Chenn and Walsh 2002). However, the brains of all vertebrates have a common structural organization, consisting of the cerebral hemispheres, diencephalon, midbrain, cerebellum, pons, and medulla (Sarnat and Netsky 1981). Among mammals, the neural structures within these divisions and

the circuits that interconnect them have extensive similarities. For mice in particular, there are extensive genetic and neuroanatomical homologies that give rise to a wide variety of behavioral processes which are well conserved between mice and humans. For instance, mice and humans both display complex behaviors, such as hunger, fear, aggression, circadian rhythms, classical and operant conditioning, and sexual behavior (Chemelli et al. 1999; Campbell et al. 1999; Hara et al. 2001). In addition, many of these behaviors show genetic differences between various inbred mouse strains (Crawley 1996; Gerlai 1996). For example, differences between common laboratory strains such as 129SvEvTac, A/J, C57BL/6J, DBA/2J, C3H/HeJ, and FVB/NJ exist for open field locomotion, learning and memory tasks, aggressive behaviors, reproductive behaviors, and acoustic startle and prepulse inhibition phenotypes (Crawley et al. 1997). Therefore, mice provide an alternative to using human populations in dissecting the genetic basis of complex behaviors which have been shown to harbor strong genetic components in humans via heritability estimates (Anokhin et al. 2003). For example, it would be possible to perform gene expression analyses on neural tissue harvested from mice varying in specific behaviors or perform QTL analysis on an F2 cross between two parental strains that significantly differ in a specific behavior analogous to a behavior of interest in humans (Chesler et al. 2005).

However, some of the same problems mentioned above with respect to human genetic studies occur when using mouse models. For example, gene expression analysis in neural tissue harvested from mice often identifies numerous genes as differentially expressed between disparate phenotypes. In addition, QTL mapping in

mice usually discovers large genomic blocks with numerous genes that require extensive follow up work (Darvasi 1998). QTL analysis in mice also necessitates several breeding steps to create useful generations, and phenotypes have to be scored at each step, which can be laborious (Silver 1995). Lastly, and maybe most importantly, certain phenotypes are not amenable to QTL analysis, specifically disease and aging phenotypes, because by the time the phenotype can be scored the mice are no longer able to breed (Johnson et al. 1992). However, the integration of gene expression data with SNP and phenotype data on inbred mouse strains offers avenues around the obstacles presented by traditional QTL and gene expression studies.

The use of inbred mouse strains to dissect the genetic complexity of common complex human diseases by integrating gene expression data with data from computationally derived association analyses provides a powerful, unbiased way to identify candidate genes controlling complex phenotypes. The endogenous genetic variation that exists among inbred mouse strains can be exploited to identify genetic control of complex physiologic processes involved in human disease (Frankel 1995; Beck et al. 2001). Recent advances in genetics and gene expression technology have greatly increased the knowledge that can be derived from this approach when applied to traditional genetic studies (Schadt et al 2005). Currently, there are more that 160,000 SNPs on over 40 inbred mouse strains (Pletcher et al. 2004; Cervino et al. 2005). This SNP information can be combined with phenotype databases, such as the Jackson Laboratory's Mouse Phenome Database, to perform association analyses (Grupe et al. 2001; Grubb et al. 2004). These results can then be integrated with gene

expression studies in multiple ways to identify high-yield candidate genes. Part of the focus of this dissertation is on gene expression patterns within the mouse brain (as detailed in Chapter 2), variation in widely used inbred mouse strains concerning gene regulation in the brain, relevant sequence differences across the strains with respect to gene regulation (as detailed in Chapter 3), and neurobehavioral phenotypic divergence. Analysis strategies have been developed to facilitate this work that attempt to integrate these very disparate data sources and gain additional biological insights (as detailed in Chapters 4, 5 and 7).

Genetical Genomics – An Example of High-Dimensional Data Integration

Recently, researchers have begun to integrate DNA sequence variation/genotype information with gene expression information in segregated populations in what has been called "genetical genomics" or expression quantitative trait locus ("eQTL") mapping studies (Schadt et al. 2003; Monks et al. 2004; Cheung et al. 2005; Bystrykh et al. 2005). These studies seek to identify DNA sequence variations that influence the expression levels of particular genes and have, in fact, shown that naturally occurring DNA sequence variation in a wide variety of organisms influences the level of expression of particular genes (Brem et al. 2005; Storey et al. 2005). It is not unexpected that DNA sequence variations, such as single nucleotide polymorphisms or deletions, in gene regulatory regions of the genome, such as promoters, could influence the ability of a transcription factor to bind. Affecting the binding ability of the transcription factor could alter the activity of the promoter in

guiding transcription of the gene. However, the ability to catalogue these variations on a genomic scale has been unimaginable until the advent of genomic technologies such as microarrays.

Many studies examining the relationship of DNA sequence variations and gene expression levels have focused on a single specific tissue and have not attempted to compare heterogeneous tissues, such as different regions of the brain, with respect to gene expression patterns and regulatory networks. Chapter 3 attempts to look at particular regions of the brain and how brain region specificity may dictate transcriptional regulatory networks on a genomic scale. Moreover, previous eQTL studies have not actually considered the biological mechanisms behind such regulatory relationships, but have rather focused on the mere association between sequence variation and gene expression patterns in an effort to make broad claims about the role of likely *cis* vs. *trans* factors in mediating gene expression. Nevertheless, these studies have shed enormous light on the global role of sequence variation in mediating gene expression (Rosa et al. 2006). However, the simple 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 is 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. It is well known that genes operate in networks replete with redundancy, feedback, and compensatory mechanisms. In fact, most traits or diseases are multifactorial and complex genetically, whereby many genes and environmental factors are responsible for their expression.

Work detailed in Chapter 7 looks at the ability to integrate gene expression information, sequence variation and biochemical pathway information to create more biologically meaningful eQTL associations. The goal of the analysis is to determine if it is possible to make sense of the collection of genes all within the same pathway whose expression patterns are influenced by a specific SNP or group of SNPs.

High-dimensional Data Analyses

A significant obstacle to the use of genomic technologies in dissecting the genetically-mediated inner workings of the brain is analyzing and integrating the vast amount of information generated by these experiments. Within the last decade, the ability to generate massive amounts of biological data has exceeded our capacity to

make sense of that data (Bassett et al. 1999). This phenomenon of data overload is particularly evident in the genomics and genetics fields where it has become routine to produce sequencing, genotyping, and gene expression data generating millions of data points (Dennis 2002). Thus, it is often the case that a researcher examining gene expression differences between mice using microarrays will produce many more data points (i.e. gene expression signals) than units of observation (i.e. individual mice assayed). This situation can create the potential for false positive results due to the number of test statistics that might be generated (Pounds and Morris 2003; Pounds and Cheng 2004), as well as the potential for false negative results if the effect induced by the intervention on gene expression involves subtle perturbations in a number of genes whose individual effects are hard to discern. Therefore appropriate analysis methodologies must be developed which consider the multivariate nature of the information produced by technologies, such as microarrays. Such analysis tools must consider error in the measurements, the definition and limitations of multivariate profiles or phenotypes arising from an appropriate data analysis, and the ability to leverage ancillary information of relevance, such as sequence data in the context of differences in the expression of a gene between two individuals. We have developed and adapted approaches described in this thesis for the analysis of gene expression data in particular that are consistent with these considerations (Chapters 5 and 6). One can consider individual gene expression values as providing insight into much larger macro-physiologic processes and mechanisms, such as regulatory pathways and

networks, whose perturbations can really only be evaluated when a number of gene expression values are considered simultaneously (Subramanian et al. 2005).

As an example consider the fact that researchers often have specific hypotheses about how their samples assayed by high-dimensional genomic technologies might be organized. In these circumstances, it is common to test whether or not sets of variables are associated with each other, or whether or not there are differences between sets of data grouped together in various ways. With large amounts of data, one could potentially test many hypotheses or group differences, creating the potential for multiple comparisons statistical issues and false positives. Therefore, multivariate statistical techniques which treat groups of data as whole systems rather than unrelated parts are appropriate and required (as detailed in Chapter 6).

In other settings, researchers have minimal insight into what variables or groups of variables might be related or worth investigating. In these situations the discovery and/or identification of non-random patterns in the data is of interest. The ultimate biological meaningfulness of identified patterns is often called into question with these analyses. An example in which researchers have no *a priori* hypotheses about how their data might be organized but instead wish to identify specific patterns within their data would be trying to determine a unique pattern in microarray-based gene expression data that appears to capture the effect of a previously unidentified biological process or phenomenon, such as which genes best discriminate between patients who will go on to develop metastatic cancer versus those who will not. Rather than utilizing traditional pattern recognition algorithms which do not use

inferential statistics to make predictions, such as hierarchical or Bayesian clustering, neural networks, principal component analysis or machine learning, we have developed multivariate statistical techniques that allow for predictor variables to be statistically tested against groups of data points to help identify maximal discriminates not due to random chance (as detailed in Chapter 5).

Lastly, it is often the case that researchers have high-dimensional data from disparate sources and their goal is to integrate this data to test specific hypotheses or to draw inferences about the biological meaningfulness of non-random patterns emerging from large data sets. The quality of the ancillary data brought to bear on the analysis of a particular data set is important if compelling claims or inferences are to be drawn. An example integration analysis setting would be an analysis of microarray-based gene expression data associated with genetic variation (i.e. SNPs) using knowledge of the biochemical pathway in which each of the genes assayed are known to participate (as detailed in Chapter 7).

There are many traditional research fields that have considered aspects of the above high-dimensional data analysis settings. For example, the growing research field of bioinformatics has, at its core, the organization and analysis of genomic data. However, much of bioinformatics is concerned with the identification of ubiquitous and universal patterns in, for example, sequence data (such as the motifs that are overrepresented in promoters). In addition, a great deal of methodology used in bioinformatics research focuses on pattern discovery using techniques developed by computer scientists, such as neural networks, machine learning or self-organizing

maps, which do not necessarily take advantage of inference-making techniques developed by mathematical statisticians. Examples of these statistical techniques would be knowledge of the asymptotic properties of test statistics derived to test hypotheses (Hastie et al. 2001). Statistical genetics has been a research field devoted to the application of extensions of traditional inference-drawing procedures to high dimensional DNA sequence variation data. But, the field has focused almost exclusively on the exploitation of the algebraic patterns of heredity (Mendel's laws) in order to identify inherited factors that may contribute to the expression of a trait or disease. Finally, traditional biostatistical analyses have focused on issues such as study design, the construction of specific low dimensional hypothesis testing settings, or generic analysis methods that can be used in a wide variety of data settings. There has been a lack of integration of these disciplines in genomics in general and specifically in neurogenomics. The body of work presented herein attempts to assimilate the distinct fields of neuroanatomy, genomics, bioinformatics, statistical genetics and biostatistics to create novel inference-drawing analysis tools and to develop new insights into brain-related biological processes.

Specific Aims

The convergence of sequence data, gene expression data and phenotypic data has been occurring on a genomic scale (Lockhart and Winzeler 2000; Grubb et al. 2004; Pletcher et al. 2004). It is now possible to analyze genomic sequence data, such as SNPs, in parallel with large scale gene expression data in what has been called a

genetical genomics approach to determine relationships to multiple well-quantified phenotypes (Schadt et al. 2003; Chesler et al. 2005; Schadt et al. 2005). This theoretically allows one to identify multiple genes working together on several related phenotypes. The ability to query these rich, complimentary data sources together offers tremendous inroads to start to unravel the genetic determinates of complex polygenic diseases and phenotypes, such as specific neuropsychiatric diseases (Carter et al. 2001). The overarching goal of the proposed research is to leverage genomic data, such as high density SNP maps and gene expression data, and develop novel analysis methods to gain insight into the functions of the brain. The specific aims of the proposed research are:

1. Create a gene expression based genomic map of the adult mammalian CNS in a model system, *Mus musculus*, as a starting point to relate gene expression, anatomy, function and phenotype (Zapala et al. 2005; Chapter 2).
2. Study the relationships between sequence variation and gene expression variation across the genome in multiple inbred mouse strains and several brain regions to identify transcriptional regulatory mechanisms within the brain on a genomic scale (Hovatta, Zapala et al., 2007; Chapter 3).
3. Develop methods that control for potential false positives in expression QTL experiments because of sequence variation that affects probe hybridization on microarrays (Greenhall, Zapala et al., 2007; Chapter 4).

4. Develop a multivariate distance matrix regression (MDMR) technique that will assist in the identification of candidate genes from both gene expression and association mapping studies (Zapala and Schork, 2006, Chapter 5) and explore the statistical properties of MDMR (Zapala and Schork, in review, Chapter 6).
5. Integrate genomic gene expression and SNP data for multiple mouse strains using sophisticated association mapping techniques that include biochemical pathway information into the association (Ye, Zapala et al., in review, Chapter 7).

References

Accili D. (2004). A note of caution on the Knockout Mouse Project. *Nat Genet*. 36:1132.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 96:6745-50.

An P, Freedman BI, Hanis CL, Chen YD, Weder AB, Schork NJ, Boerwinkle E, Province MA, Hsiung CA, Wu X, Quertermous T, Rao DC. (2005). Genome-wide linkage scans for fasting glucose, insulin, and insulin resistance in the National Heart, Lung, and Blood Institute Family Blood Pressure Program: evidence of linkages to chromosome 7q36 and 19q13 from meta-analysis. *Diabetes*. 54:909-14.

Anokhin AP, Heath AC, Myers E, Ralano A, Wood S. (2003).Genetic influences on prepulse inhibition of startle reflex in humans. *Neurosci Lett*. 353:45-8.

Austin CP, Battey JF, Bradley A, Bucan M, Capecchi M, Collins FS, Dove WF, Duyk G, Dymecki S, Eppig JT, Grieder FB, Heintz N, Hicks G, Insel TR, Joyner A, Koller BH, Lloyd KC, Magnuson T, Moore MW, Nagy A, Pollock JD, Roses AD, Sands AT, Seed B, Skarnes WC, Snoddy J, Soriano P, Stewart DJ, Stewart F, Stillman B, Varmus H, Varticovski L, Verma IM, Vogt TF, von Melchner H, Witkowski J, Woychik RP, Wurst W, Yancopoulos GD, Young SG, Zambrowicz B. (2004). The knockout mouse project. *Nat Genet*. 36:921-4.

Barlow C, Lockhart DJ. (2002). DNA arrays and neurobiology--what's new and what's next? *Curr Opin Neurobiol*. 12:554-61.

Bassett DE, Eisen MB, Boguski MS. (1999). Gene expression informatics--it's all in your mine. *Nat Genet*. 21:51-5

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM. (2000). Genealogies of mouse inbred strains. *Nat Genet*. 24:23-5.

Boffelli D, Nobrega MA, Rubin EM. (2004). Comparative genomics at the vertebrate extremes. *Nat Rev Genet*. 5:456-65.

Botstein D, White RL, Skolnick M, Davis RW. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980 32:314-31.

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

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet*. 37:225-32.

Campbell KM, de Lecea L, Severynse DM, Caron MG, McGrath MJ, Sparber SB, Sun LY, Burton FH. (1999). OCD-like behaviors caused by a neuropotentiating transgene targeted to cortical and limbic D1+ neurons. *J Neurosci*. 19:5044-53.

Carter TA, Del Rio JA, Greenhall JA, Latronica ML, Lockhart DJ, Barlow C. (2001). Chipping away at complex behavior: transcriptome/phenotype correlations in the mouse brain. *Physiol Behav*. 73:849-57.

Cervino AC, Li G, Edwards S, Zhu J, Laurie C, Tokiwa G, Lum PY, Wang S, Castellini LW, Lusis AJ, Carlson S, Sachs AB, Schadt EE. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. *Genomics*. 86:505-17.

Chemelli RM, Willie JT, Sinton CM, Elmquist JK, Scammell T, Lee C, Richardson JA, Williams SC, Xiong Y, Kisanuki Y, Fitch TE, Nakazato M, Hammer RE, Saper CB, Yanagisawa M. (1999). Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. *Cell*. 98:437-51.

Chenn A, Walsh CA. (2002). Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science*. 297:365-9.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*. 37:233-42.

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.

Crawley JN. (1996). Unusual behavioral phenotypes of inbred mouse strains. *Trends Neurosci*.19:181-2.

Crawley JN, Belknap JK, Collins A, Crabbe JC, Frankel W, Henderson N, Hitzemann RJ, Maxson SC, Miner LL, Silva AJ, Wehner JM, Wynshaw-Boris A, Paylor R. (1997). Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology*. 132:107-24.

Darvasi A. (1998). Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet*. 18:19-24.

Dennis C. (2002). Information overload. *Nature*. 417:14.

Frankel WN. (1995). Taking stock of complex trait genetics in mice. *Trends Genet*. 11:471-7.

Gerlai R. (1996). Gene-targeting studies of mammalian behavior: is it the mutation or the background genotype? *Trends Neurosci*. 19:177-81.

Geyer MA, McIlwain KL, Paylor R. (2002). Mouse genetic models for prepulse inhibition: An early review. *Molecular Psychiatry* 7:1039-53.

Greenhall JA*, Zapala MA*, Caceres M, Libiger O, Barlow C, Schork NJ, Lockhart DJ. (2006). Mining array-based gene expression data for evidence of genetic variation. *Genome Research*. In Press.

Grubb SC, Churchill GA, Bogue MA. (2004). A collaborative database of inbred mouse strain characteristics. *Bioinformatics* 20:2857-9.

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. (2001). In silico mapping of complex disease-related traits in mice. *Science*. 292:1915-8.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*. 306:234-8.

Hara J, Beuckmann CT, Nambu T, Willie JT, Chemelli RM, Sinton CM, Sugiyama F, Yagami K, Goto K, Yanagisawa M, Sakurai T (2001). Genetic ablation of orexin neurons in mice results in narcolepsy, hypophagia, and obesity. *Neuron*. 30:345-54.

Hastie T, Tibshirani R, Friedman JH. (2001). *Elements of Statistical Learning*. Springer.

Hovatta I*, Zapala MA*, Broide RS, Schadt EE, Schork NJ, Lockhart DJ, Barlow C. (2007). Relationships between inbred mouse strains based on global gene expression patterns do not correlate with known genealogy or DNA-level variation. *Genome Biology*. In press.

Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, Redwine JM, Ellison JA, Schadt EE, Verma IM, Lockhart DJ, Barlow C. (2005). Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature*. 438:662-6.

Hyman SE. (2002). Neuroscience, genetics, and the future of psychiatric diagnosis. *Psychopathology*. 35:139-44.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*. 409:860-921.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*. 437:1299-320.

Johnson TE, DeFries JC, Markel PD. (1992). Mapping quantitative trait loci for behavioral traits in the mouse. *Behav Genet*. 22:635-53.

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*.445:168-76.

Lombardino AJ, Hertel M, Li XC, Haripal B, Martin-Harris L, Pariser E, Nottebohm F. (2006). Expression profiling of intermingled long-range projection neurons harvested by laser capture microdissection. *J Neurosci Methods*. 157:195-207.

Lockhart DJ, Winzeler EA. (2000). Genomics, gene expression and DNA arrays. *Nature*. 405:827-36.

Macauley A, Ladiges WC. (2005). Approaches to determine clinical significance of genetic variants. *Mutat Res*. 573:205-20.

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.

Nadler JJ, Zou F, Huang H, Moy SS, Lauder J, Crawley JN, Threadgill DW, Wright FA, Magnuson TR. (2006). Large-scale gene expression differences across brain regions and inbred strains correlate with a behavioral phenotype. *Genetics*. 174:1229-36.

Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW, Lagler E, Korstanje R, Wang X, Nusskern D, Bogue MA, Mural RJ, Paigen B, Wiltshire T. (2004). Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol.* 2:e393.

Pounds S, Morris SW. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*. 19:1236-42.

Pounds S, Cheng C. (2004). Improving false discovery rate estimation. *Bioinformatics*. 20:1737-45.

Reeders ST, Breuning MH, Davies KE, Nicholls RD, Jarman AP, Higgs DR, Pearson PL, Weatherall DJ. (1985). A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature*. 317:542.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*. 245:1066-73.

Rosa GJ, de Leon N, Rosa AJ. (2006). Review of microarray experimental design strategies for genetical genomics studies. *Physiol Genomics*. 28:15-23.

Sarnat HB, Netsky MG: Evolution of the Nervous System. New York, Oxford University Press, 1981.

Schork NJ. (2002). Power calculations for genetic association studies using estimated probability distributions. *Am J Hum Genet*. 70:1480-9.

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.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 37:710-7.

Schadt EE. (2005). Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr Opin Biotechnol*. 16:647-54.

Silver LM. (1995). Classic linkage analysis and mapping panels. in *Mouse Genetics: Concepts and Applications*. Oxford University Press, Oxford, UK.

Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, O'Brien WE, Conti DV, Witte JS, Lander ES, Nadeau JH. (2004). Genetic dissection of complex traits with chromosome substitution strains of mice. *Science*. 304:445-8.

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

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 102:15545-50.

Tecott LH. (2003). The genes and brains of mice and men. *Am J Psychiatry*.160:646-56.

Wade CM, Daly MJ. (2005). Genetic variation in laboratory mice. *Nat Genet*. 37:1175-80.

Zapala MA, Hovatta I, Ellison JA, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Lockhart DJ, Reilly JF, Young WG, Bloom FE, Lockhart DJ, Barlow C. (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*. 102:10357-62.

Zapala MA, Barlow C, Hovatta I. (2007). Molecular Anatomy of the Mammalian Brain. In F. Bloom, F. Gage & L. Squire (Eds.). *New Encyclopedia of Neuroscience*. New York, Elsevier. In Press.

Zhang Z, Carriero N, Gerstein M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet*. 20:62-7.

CHAPTER 2

Adult Mouse Brain Gene Expression Patterns Bear an Embryologic Imprint

Abstract

The current model to explain the organization of the mammalian nervous system is based on studies of anatomy, embryology, and evolution. To further investigate the molecular organization of the adult mammalian brain, we have built a gene expression-based brain map. We measured gene expression patterns for 24 neural tissues covering the mouse central nervous system and found, surprisingly, that the adult brain bears a transcriptional "imprint" consistent with both embryological origins and classic evolutionary relationships. Embryonic cellular position along the anterior-posterior axis of the neural tube was shown to be closely associated with, and may be a determinant of, the gene expression patterns in adult structures. We also observed a significant number of embryonic patterning and homeobox genes with region-specific expression in the adult nervous system. The relationships between global expression patterns for different anatomical regions and the nature of the observed region-specific genes suggest that the adult brain retains a degree of overall gene expression established during embryogenesis that is important for regional specificity and the functional relationships between regions in the adult. The complete collection of extensively annotated gene expression data along with new data mining and visualization tools have been made available on a publicly accessible website (http://www.barlow-lockhartbrainmapnimhgrant.org/).

Introduction

The adult nervous system achieves its mature form as the result of neuroectodermal cells committing to a unique fate and then segregating into distinct regional collectives of neurons which become fully functional through establishment of connections to other neurons. Our current understanding of brain architecture and organization is based on studies of embryology, anatomy and evolution in which direct observation of anatomic structures was the foundation for postulated models of brain structure (Swanson 2000). Recent models of brain development and maturation consider relationships between different regions based on the expression of specific genes in assigning developmental origins of adult structures (Holland and Holland 1999; Puelles and Rubenstein 2003). Here, we have constructed a regional gene expression atlas of the adult mouse brain and analyzed the quantitative results using molecular classification algorithms.

Genome-wide gene expression profiling is a powerful technique for deriving information about specific brain regions (Lockhart and Barlow 2001; Barlow and Lockhart 2002). This approach has been used to measure gene expression patterns in particular regions, subregions or cell populations in the brain (Zhao et al. 2001; Zirlinger et al. 2001; Bonaventure et al. 2002; Kamme et al. 2003; Zirlinger et al. 2003; Lein et al. 2004). Two previous studies have analyzed gene expression differences across multiple regions of the mammalian brain using multiple strains or species (Sandberg et al. 2000; Khaitovich et al. 2004). However, the current study is the most extensive to date in terms of the number of genes and the coverage of

different neural tissues. Our goal was to create a publicly accessible gene-based brain map, with datasets, metadata, databases and analysis tools available for use by the scientific community (Barlow and Lockhart 2002). As part of this work, we measured gene expression patterns for 24 neural tissues covering the adult mouse central nervous system plus 10 body regions. The gene expression data along with new data mining and visualization tools are available on a publicly accessible website (http://www.barlow-lockhartbrainmapnimhgrant.org/). A large-scale, systematic, quantitative mouse brain gene expression database, called TeraGenomics, was built to house and provide access to all the quantitative, region-specific gene expression data, along with quality control measures, anatomical information, strain information, dissection protocols, sample preparation information, and array hybridization parameters in accordance with MIAME (Minimal Information About a Microarray Experiment) (Brazma et al. 2001).

Our goal in this study is to understand how regional gene expression patterns in the brain are related to brain architecture and organization. We sought to identify relationships between brain regions based on both shared and restricted gene expression patterns. The gene expression data were analyzed using molecular classification algorithms, without pre-specified anatomical information, to define relationships between brain structures. To our surprise, we found that the gene expression patterns of the adult brain have a transcriptional "imprint" that is consistent with embryological origins and classic evolutionary relationships between subregions of the cortex.

Materials and Methods

Tissue Collection.

All animal procedures were performed according to protocols approved by The Salk Institute for Biological Studies and BrainCells Inc. Animal Care and Use Committees. Male A/J, C57BL/6J (B6), C3H/HeJ and DBA/2J (DBA) mice were purchased from The Jackson Laboratories (Bar Harbor, ME); male 129S6/SvEvTac (129) mice were purchased from Taconic Farms, Inc. (Germantown, NY). All mice were purchased at an age of 7 weeks and housed individually for 1 week before sacrifice. Dissections were done between 11.00-17.00 h. Mice were sacrificed by either cervical dislocation or decapitation and dissected tissue, collected within 15 minutes of sacrifice, was directly frozen on dry ice and stored at -80°C. The following brain regions were collected: amygdala (Amg), bed nucleus of the stria terminalis (Bnst), CA1 region of the hippocampus (CA1), CA3 region of the hippocampus (CA3), cerebellum (Cb), choroid plexus from the 4th ventricle (cp4v), neocortex (Cx), dentate gyrus (DG), diencephalon & midbrain, excluding hypothalamus (DiE-MD), entorhinal cortex (EntCx), hippocampal formation (HiF), hypothalamus (Hy), inferior colliculus (IC), medulla oblongata (MO), motor cortex (MtrCx), olfactory bulbs (Olf), periaqueductal gray (Pag), perirhinal cortex (PrhCx), pituitary (Pit), pons (Pons), retina (Retina), spinal cord (SpCrd), striatum (Str), and superior colliculus (SC). The following body regions were collected: adrenal glands, brown adipose tissue (retroperitoneal and interscapular), heart, kidney, liver, skeletal muscle (femoral),

spleen, testes, thymus, white adipose tissue (epididymal) (for a description of the samples see Table 1 which is published as supporting information on the PNAS web site).

To ensure that highly reproducible dissections were conducted for each region, bregma coordinates and anatomic boundaries defining each region were established based upon the Paxinos and Franklin mouse brain atlas (Paxinos and Franklin 2001). A reference document was created consisting of photographs and atlas bregma coordinates to illustrate the exact methods used to dissect each region, including step-by-step instructions (for an example, see Figure 3 which is published as supporting information on the PNAS web site). The dissection reference documents accompany the processed microarray data as part of the MIAME (Brazma et al. 2001) compliant metadata housed in the publicly accessible relational database (http://www.barlow-lockhartbrainmapnimhgrant.org/). The metadata contains 75 different fields of sample annotation which include dissection protocols and anatomical information defining the bregma coordinates (all dissection protocols and metadata are available at http://www.barlow-lockhartbrainmapnimhgrant.org/, examples are available as pdf files at the PNAS web site). In addition, the anatomical hierarchy of the Neuronames taxonomy (Bowden and Martin 1995) has been included as a user-friendly query tool within the database.

RNA Preparation.

Total RNA was isolated according to the methods published in Sandberg et al.

(Sandberg et al. 2000). Tissues were placed into TRIzol (GIBCO/BRL) (added to the frozen tissues approximately 1 ml per 100 mg tissue) and homogenized (Polytron, Kinematica, Lucerne, Switzerland) at maximum speed for 90-120 seconds. Subsequent steps were performed according to the manufacturer's instructions for all tissues with the exception of spleen and white adipose in which the Qiagen RNeasy Mini Kit was used to clean up the total RNA after the TRIzol protocol. White adipose RNA was prepared using a protocol kindly provided by Eric Muise and Yarek Hrywna of Merck Research Labs, Rahway, NJ. Tissues were added to 4 ml TRIzol and homogenized for 90 seconds. Following a ten minute incubation at room temperature, samples were spun for ten minutes at 3,200 g and the top fat layer that resulted was removed. After the addition of chloroform, the samples were spun for 20 minutes at 3,200 g. The rest of the protocol was performed according to the TRIzol instructions, and the Qiagen RNeasy Mini Kit was used to clean up the total RNA. Labeling of all samples, hybridization, and scanning were performed using a modification of the protocol developed by Wodicka et al. (Wodicka et al. 1997) using the Affymetrix Genechip MG_U74Av2 microarray that contains 12,422 probe sets corresponding to approximately 12,000 genes and expressed sequence tags (Affymetrix, Santa Clara, CA).

Database and Analysis Tools.

After scanning the arrays with the Affymetrix GeneArray Scanner, the .cel files were uploaded, housed and analyzed in the Teradata analytical relational database

(Teradata, a divison of NCR, Dayton, OH) with algorithms developed by our laboratory using the TeraGenomics software tool [Information Management Consultants, Inc., Reston, VA] (Caceres et al. 2003). Additional analysis was performed using the freeware program Bullfrog version 10.2 [(Zapala et al. 2002) for original version, current version of Bullfrog is available at http://www.barlow-lockhartbrainmapnimhgrant.org/], Genespring v6 (Silicon Genetics, Redwood City, CA), the Gene Ontology Tree Machine [http://genereg.ornl.gov/gotm/ (Zhang et al. 2004)] and Excel. The signals for each array were scaled to an overall target intensity of 200 (Wodicka et al. 1997) and arrays were normalized separately to the same average intensity based on the probe sets corresponding to the 60^{th} to 90^{th} percentile of hybridization signals.

Analysis algorithms and criteria.

The algorithms and criteria used to analyze the gene expression data in TeraGenomics are described in *Supporting Methods*, which is published as supporting information on the PNAS web site.

Microarray quality, experimental reproducibility and false positives.

Given the large amount of data in the atlas, the quality of the samples were assessed at several steps including total RNA quality (minimum yield was 10 μg with 260/280 ratio in Tris-EDTA between 2.0-2.2), cRNA yield and quality (minimum cRNA yield was 0.66 μg/μl), array hybridization quality control metrics (for all arrays

background was < 200; raw Q < 5; scaling factor < 6; outliers < 500; percent present or marginal ≥ 45%; actin 3'/5' < 2; and GAPDH 3'/5' < 2), and by assessing the performance of replicates (Pearson correlation coefficient between replicates required to be ≥ 0.97 and the number of genes scored as different between replicates required to be less than 1% of the total number of probe sets on the arrays) (see Table 2 which is published as supporting information on the PNAS web site). To determine experimental reproducibility and false positive rates, we compared independent samples from different animals from the same region and same strain. The set of criteria used to establish experimental reproducibility between replicates was a fold change of 1.5 or greater, a difference call of increase, marginal increase, decrease or marginal decrease and a signal change (scaled) greater than 30 in 100% of comparisons. Since these comparisons were between replicate groups, by definition any genes returned as significantly different would be considered false positives (see Table 3 which is published as supporting information on the PNAS web site).

Heat map and Cluster Analysis.

A correlation matrix of brain region relatedness was generated for all 100 pairwise comparisons. From this analysis we observed that the average intrastrain replicate R value (all regions) was 0.988 for B6, 0.987 for 129 and 0.978 for DBA. The average interstrain replicate R value was 0.974 for B6 versus 129, 0.961 for B6 versus DBA and 0.963 for 129 versus DBA. Given the strong similarity between the intrastrain and interstrain comparisons within a particular brain region, for the

purposes of this study we averaged the data for each brain region independent of strain. The Bullfrog software was used to identify the 7852 probe sets that were called "Present" (P) and with a scaled signal of 35 or greater in at least one of the 24 neural tissues. The 'most variable' genes from this subset were then identified using an algorithm that normalized signal values across the 24 regions for each gene and ranked the genes from highest to lowest using the standard deviation of the normalized signals. The probe sets with a standard deviation greater than 0.15 were identified yielding a total of 4894 genes. Matlab Student 7.0 was used to generate a heat map and Genespring v6 was used to generate the clustering relationship based on the Pearson correlation for all pairwise comparisons of absolute signal intensity (Figure 1, and Table 4 which is published as supporting information on the PNAS web site).

Identification of Region Restricted or Region Enriched Gene Expression Patterns.

To identify genes with region restricted or region enriched expression patterns, probe sets that were called P and with a signal of 35 or greater in at least one sample were used in the analysis (8156 probe sets). Data were analyzed for 22 mouse brain regions (excluding cp4v and Pit) to identify genes that are clearly expressed at detectable levels in only one to two distinct brain regions. Data from two different inbred mouse strains (two replicates per strain) were analyzed for each of the brain regions except retina. For retina, four samples from one strain were used in this analysis. Data files were exported from TeraGenomics and a combination of filtering

and ‘Venn’ functions were used in Bullfrog to identify region-restricted genes. Probe sets that were consistently detected as present in both strains for only one to two specific brain regions and consistently NOT detected in most other brain regions were identified. For genes to be categorized as region-specific or region-restricted, probe sets corresponding to those genes were required to meet a set of empirically-derived selection criteria which were based on the “Present” and “difference” calls (see the selection criteria described in *Supporting Methods*, which is published as supporting information on the PNAS web site).

To allow searches for user-defined gene expression patterns, we developed algorithms in Bullfrog to identify genes enriched in specific regions. For this purpose, the normalized signal intensities of the replicate samples were averaged. The ‘shape vector’ analysis tool in Bullfrog was used to identify probe sets with expression ‘vectors’ [normalized signals across the 23 brain regions (excluding cp4v)] that were most highly correlated with an entered ‘ideal’ pattern. For example, to find genes specifically enriched in the Amg, the following ‘ideal shape’ vector was used: (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) for the regions Amg, Bnst, CA1, CA3, Cb, Cx, DG, EntCx, HiF, Hy, IC, DiE-MD, MO, MtrCx, Olf, Pag, PrhCx, Pit, Pons, Retina, SpCrd, Str, SC. The data were then sorted based on the correlation coefficient (R) between the observed and the ideal pattern, to yield the list of genes with expression patterns that match the input pattern most closely. The enriched genes are shown in Table 6 which is published as supporting information on the PNAS web site.

Gene Ontology Analysis.

The web-based program GOTree Machine (GOTM) was used for the gene ontology analysis [http://genereg.ornl.gov/gotm/ (20)]. Excluding the genes specific for retina, Pit and cp4v, 192 region-restricted and enriched genes were analyzed using the GOTM (see Table 7 which is published as supporting information on the PNAS web site for a list of the genes). GOTM was used to identify gene ontology (GO) categories with representations significantly different than expected by chance ($p < 0.01$). This analysis was carried out October 2004.

Digital brain atlas.

The digital atlas was generated from a 90 day-old C57BL/6N mouse brain that was prepared as follows. C57BL/6N (Charles River, MA) mice were anesthetized with Avertin (0.5 mg/g body weight, i.p.) followed by transcardial perfusion with a light (10%) sucrose solution. The brains were removed, immediately frozen in isopentane at -30ºC and stored at -80ºC until use. The brain chosen for the atlas was cryostat-cut (30 μm-thick) in the coronal plane of section over a 2 day period resulting in a total of 462 sections mounted onto 110 microscope slides. The sections, which span from the tip of the olfactory bulbs to the end of the cerebellum, were Nissl-stained with a combination of Cresyl Violet and Thionin, providing enhanced differentiation between neurons and glia. Brain sections for the atlas were then digitally acquired into a database, using in-house, JAVA-based software, NeuroZoom^{TM}, at a very high

resolution of 1.3 μm/pixel. The individual image tiles were then stitched together by the software and the resulting 462 files take up a total of 52 Gb of memory.

In order for the atlas to be visualized in arbitrary planes of section and in a 3-dimensional (3D) virtual display (Figure 2C), the digitized brain sections were aligned, and are in register with, a magnetic resonance microscopy (MRM) data file collected at 11.7 T for a formaldehyde fixed C57BL/6N brain within the skull. The atlas sections were synchronized by surface alignments to the MRM file using center-alignment algorithms to register the atlas section contours to those from the MRM file. This MRM file is one of 10 similar MRM datasets for 90 day-old C57BL/6N male mice that we have captured and is a true representation of a C57BL/6N brain with minimal inter-individual variance (Redwine et al. 2003). The alignment allows the atlas to be viewed, not only in the coronal plane of section in which it was generated, but also in the extrapolated sagittal and horizontal planes, which are dynamically constructed from slices of the coronal sections, with accuracy as well as orthogonal views with rotation.

Graphical delineations of brain regions generated using the NeuroZoom™ software are closed polygons that are overlaid on top of the coronal sections. For this study, 21 major and minor regions throughout the brain, matching the regions dissected to obtain the mRNA samples, were used for display, including Amg, Bnst, HiF, Cb, Cx, Str, SC and IC, MO, Olf, Pag and SpCrd (Figure 2D). The 2-dimensional annotations of these regions were 3-dimensionally reconstructed using a surface triangulation algorithm. Data containing signal intensity values from gene expression

microarray analyses were imported by the software and converted to a linear color scale with a high-to-low gradient ranging from red, orange, yellow, green, cyan, blue, indigo, to violet (ROYGBIV) (Figure 2E). The 2D and 3D contours were filled with either a user-specified color, or with a color corresponding to a value from the color scale representing the signal intensity. In some cases, the transparency of the color-filled contours was adjusted using a scale from 0-100% transparent.

Results

Molecular Architecture of the Adult Brain.

We have profiled gene expression patterns of 24 neural tissues and 10 body regions. In total, 150 array hybridization measurements were included in our dataset. For a summary of the microarray data included in the atlas, the average quality control metrics by sample type and the experimental reproducibility, see Tables 2 and 3 which are published as supporting information on the PNAS web site.

A 'heat map' was generated to look for similarities and differences in regional gene expression patterns based on the Pearson correlation coefficients calculated between all individual samples (Figure 1A). Replicate samples from a given brain region showed the most similar gene expression profiles of all the sample groups as demonstrated by the dark red diagonal line in Figure 1A, indicating the high level of reproducibility between independent replicate measurements.

Within the cortical subregions, three groups showed very similar gene expression patterns. Expression patterns for the brain regions that comprise the

archicortex (A: CA1, CA3, DG), paleocortex (P: Amg, EntCx, PrhCx) and neocortex (N: Cx, MtrCx) were the most similar within their respective groups. Two other groups that showed very similar gene expression patterns include the Hy, Pag, IC, SC, and DiE-MD; and the Pons, MO, and SpCrd. Gene expression profiles of hindbrain regions (Pons, MO, SpCrd, Cb) were somewhat dissimilar to the profiles for structures of the forebrain and midbrain. We also noted that the patterns for three structures that develop as outpouchings of the brain (retina, pituitary and choroid plexus) were remarkably different not only from other brain structures, but also from each other. We found more similar gene expression patterns for regions that collectively shared a developmental origin, for example, DiE-MD brain structures (DiE-MD, Hy, Pag, IC, SC). These results demonstrate that position along the anterior-posterior axis of the neural tube is closely associated with, and may be a determinant of, the gene expression patterns in the adult structures.

In order to further explore the relationships between brain regions based on their gene expression profiles, unsupervised hierarchical clustering was performed (Figure 1B). We hypothesized that clustering analysis might reveal brain region relatedness based on anatomy, embryology or evolutionary relationships. The resulting dendrogram consisted of two main branches with the telencephalic brain regions on one main branch and the remaining regions clustered together on the second main branch. The first observation was that regions with shared cytoarchitectural features did not cluster together. The laminated structures (Olf, HiF, Cb, EntCx, PrhCx, MtrCx, Cx, Retina) were found on all branches of the dendrogram and were not more

similar to each other than to other regions. We next compared the branch pattern of the dendrogram to the structures of a five vesicle embryo. The majority of regions clustered together based upon the embryologic region from which they were derived (Figure 1B), demonstrating an overall region relatedness consistent with a classically defined, morphology-based embryological origin. We also noted a general preservation of the rostral-caudal axis suggested by the pattern of the heat map and the subdivisions of the dendrogram, where for example the Neocortex is more related to the Paleocortex than to the Archicortex. It is important to emphasize that the observed relationships between brain regions based on expression patterns were robust and were not significantly influenced by the particular choice of genes or the strain of mouse used for the analysis.

Region Restricted Gene Expression in the Adult Brain.

To further investigate the embryological basis for the observed region-related gene expression patterns in the adult mouse brain, we focused on defining patterns that uniquely mark a particular region or set of structures. We used a set of analyses to identify genes with highly restricted expression patterns (Sandberg et al. 2000). In one analysis, we identified 93 genes that showed expression restricted to a region or specific subregions (see Table 5 which is published as supporting information on the PNAS web site), and in a separate analysis we identified 129 genes that showed clear regional enrichment (see Table 6 which is published as supporting information on the PNAS web site), yielding 192 unique genes in total. We hypothesized that these genes

may perform functions related to regional specialization. Using the Gene Ontology Tree Machine program (GOTM) (Zhang et al. 2004), we queried the set of region specific and enriched genes (omitting the genes restricted to the Retina, Pit and cp4v, for the list see Table 7 which is published as supporting information on the PNAS web site) to identify Gene Ontology (GO) categories that were significantly over-represented ($p < 0.01$). In the biological process category, genes were over-represented for both 'development' and 'regulation of biological process' (Figure 2A). Within these two categories we found an over-representation of genes involved in morphogenesis ($p < 2.3x10^{-6}$), pattern specification ($p < 1.43x10^{-6}$), and cell communication ($p < 0.00025$) (Figure 2B). Thus, consistent with the embryonic imprint observed in the dendrogram, the GO categories for development, morphogenesis and pattern specification were over-represented in the list of region-specific genes.

In particular we observed a significant number of embryonic patterning and homeobox genes (e.g., *Dlx6*, *Gbx2*, *Chrd*, *HoxA4*, *HoxB5*) with region-specific expression in the adult nervous system. 21 out of the 192 region-specific genes were embryonic patterning genes. In studies of this type, with very large amounts of data, it is helpful to be able to visualize the data in a meaningful way. We and others have discussed the importance of methods to view data in three dimensions (3D), in the context of anatomy and/or brain circuitry (Barlow and Lockhart 2002). As a step towards this goal, we have taken the observations of embryonic patterning and homeobox gene expression in the adult brain and combined the quantitative expression

data with a high resolution, coordinate-based brain atlas that allowed us to visualize the gene expression relationships in the context of the whole brain rather than simply as a list of genes. These gene expression data were imported and visualized onto a 3D brain atlas representation (Figure 2C) using NeuroZoom^{TM} software (Neurome, Inc., La Jolla, CA) to provide a virtual *in situ* hybridization in which gene expression levels for specific brain regions are color coded. Using this display technology, we observed the expression of these embryonic patterning and homeobox genes throughout the neuraxis (Figure 2D, E). Other groups have proposed different methods to create three-dimensional brain gene expression maps, such as voxelation, where RNA is extracted from spatially registered cubes within the brain (Brown et al. 2002). Our method differs from voxelation in that we used both standard stereotaxic coordinates and strict dissection protocols that respect natural boundaries, such that the gene expression patterns represent whole subregions rather than a mixture of multiple regions. Therefore, it is not necessary to use statistical methods such as singular value decomposition (SVD) to delineate anatomical regions. With the 3D brain atlas presented here, the gene expression patterns of specific brain regions may be viewed alone, as a network with other specific regions, or in the context of the whole brain.

Discussion

Previous studies have shown that commitment to a specific lineage, specified in large part by anatomical position within the developing neural tube, involves the imprinting of a genetic program (Lumsden and Krumlauf 1996). Our expression data

suggest that the imprinted genetic program is still evident in the mature brain. The concept of an imprinted genetic pattern has been strengthened by the identification of genes that mark morphogenetic fields during brain development (Spemann 1918; Spemann and Mangold 1924). The pattern of gene expression for a small set of genes for a particular brain region and its relatedness to patterns seen in other regions has been used extensively in developmental biology to help understand the embryologic origins and functional relationships between brain regions (Holland and Holland 1999; Shimamura et al. 1995). Analysis of the relationship between morphologically defined boundaries in brain development and domains defined by gene expression patterns has led to the identification of three major regions; the anterior region, midbrain-hindbrain boundary and the rhombospinal region (Swanson 2000). The anterior region corresponds to the telencephalon, diencephalon and anterior mesencephalon, the midbrain-hindbrain boundary is the origin of the cerebellum, and the rhombospinal region corresponds to the posterior mesencephalon, metencephalon, myelencephalon, and spinal cord (Swanson 2000).

While the gene expression dendrogram observed in Figure 1B did not have three main branches corresponding to these divisions, we observed two features of the dendrogram that were consistent with the three region model. First, we observed that the cerebellum is on a branch distinct from the regions derived from the anterior or rhombospinal regions. Previously, the cerebellum was believed to be derived from the hindbrain structures along with the pons and medulla. However, it has recently been shown that the cerebellum is derived from the cells that meet at the midbrain-

hindbrain junction (Wingate 2001). The clustering results are consistent with the findings that the cerebellum is not derived from the hindbrain. Nevertheless, the cerebellum is still more closely related to brainstem structures than non-brainstem structures. It is also possible that the adult gene expression patterns of the cerebellum are so highly modified that they obscure the structure's developmental origins. Second, the brain structures comprising the rhombospinal region (Pons, MO, SpCrd) clustered together based on a high degree of expression pattern similarity. Notably, in a previous regional analysis of the amygdala, gene expression patterns in specific amygdaloid nuclei were found to respect the ontogenetic origins of the subnuclei which derive embryologically from both pallial and subpallial structures (Zirlinger et al. 2001). Like the amygdala, the BNST is known to be a heterogenous structure, and in the embryo the posterior BNST occupies a wedge between the basal ganglia and the diencephalon (Bayer 1987). The neuroepithelium from which the posterior BNST is derived lies lateral to where the anterior thalamus fuses with the hypothalamic portion of the third ventricle (Altman and Bayer 1986). This embryonic relationship between the BNST and the diencephalon, specifically the hypothalamus, appears to be observed in the gene expression patterns of the adult as demonstrated by the dendrogram (see Figure 1 B). These results suggest that while the expression pattern for many genes may change dramatically during development, the brain retains a degree of gene expression patterning established during embryogenesis that is important for maintaining regional specificity and functional relationships between brain regions in the adult.

The embryonic patterning and homeobox genes were found to be expressed in the adult brain with patterns that respected the domains and boundaries defined by the embryologic, gene expression, and classic evolutionary models of brain development and maturation, however the evolutionary models remain controversial (Figure 2D, E) (Jarvis et al. 2005). Several studies of the developing brain have demonstrated that similar sets of genes are used to establish a particular anatomical region and to maintain the cell-cell relationships of the differentiated region (Pasini and Wilkinson 2002). Thus, it may be that the roles of these genes in adulthood are similar to their roles during development. These roles include maintaining established phenotypes and connectivity of neuronal populations, or preserving barriers to the inappropriate migration of neurons from one region to another. We speculate that these genes continue to play an important role in the regional specification of functional units in the adult brain.

The expression results and the analytical and visualization tools described here add to the expanding neurobiology tool chest, and complement efforts to measure qualitative patterns of gene expression based on *in situ* hybridization (Mouse Brain Gene Expression Database project), reporter lines (Gong et al. 2003) and proteomics methods (Human Brain Proteome Project).

Acknowledgements

The authors would like to thank Dan Lockhart for creation and development of the BullFrog software; Information Management Consultants (Reston, VA) for donation of the Teradata data warehouse, design and programming of the

TeraGenomics database; Teradata/NCR (Rancho Bernardo, CA) for early support of the project; Neurome, Inc. (La Jolla, CA) for donation of time and resources; Todd Carter for technical support, Selena Ellis-Vizcarra and Jamie Simon for technical assistance; David Anderson and Roland Stoughton for helpful discussions; and Jim Velier for insights. This work was supported by grant NS039601-04 to C.B. and MH062344-03 from the National Institute of Mental Health to C.B. and D.J.L. This chapter appears as a reformatted version of the following published material:

Zapala MA*, Hovatta I*, Ellison JA*, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Reilly JF, Young WG, Bloom FE, Lockhart DJ, Barlow C. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*. 2005. 102:10357-62.

Figure 1. Heat Map and Hierarchical Cluster of Brain Map. The adult brain bears a gene expression imprint based on embryologic origin and classic evolutionary complexity. (*A*) Pearson correlation 'heat map' matrix of all brain samples. The white boxes outline the classic evolutionarily related regions of the Archicortex (A: HiF, CA1, CA3, DG), Paleocortex (P: Amg, EntCx, PrhCx), and Neocortex (N: Cx, MtrCx). Samples with very similar gene expression profiles corresponding to a higher correlation coefficient are denoted by dark red and map positions corresponding to brain regions with dissimilar gene expression profiles appear dark blue. (*B*) Unsupervised hierarchical cluster dendrogram. Shown on the left is the dendrogram relating structures to one another. Shown on the right is a schematic of the developing mouse brain with the five vesicle regions color coded on a schematic of the developing embryo. The color chart to the left shows the derivatives of these embryonic brain vesicles in the context of the dendrogram. The hatched boxes indicate brain structures formed by inductive events. A is Archicortex; P is Paleocortex; N is Neocortex.

Figure 1. Heat Map and Hierarchical Cluster of Brain Map

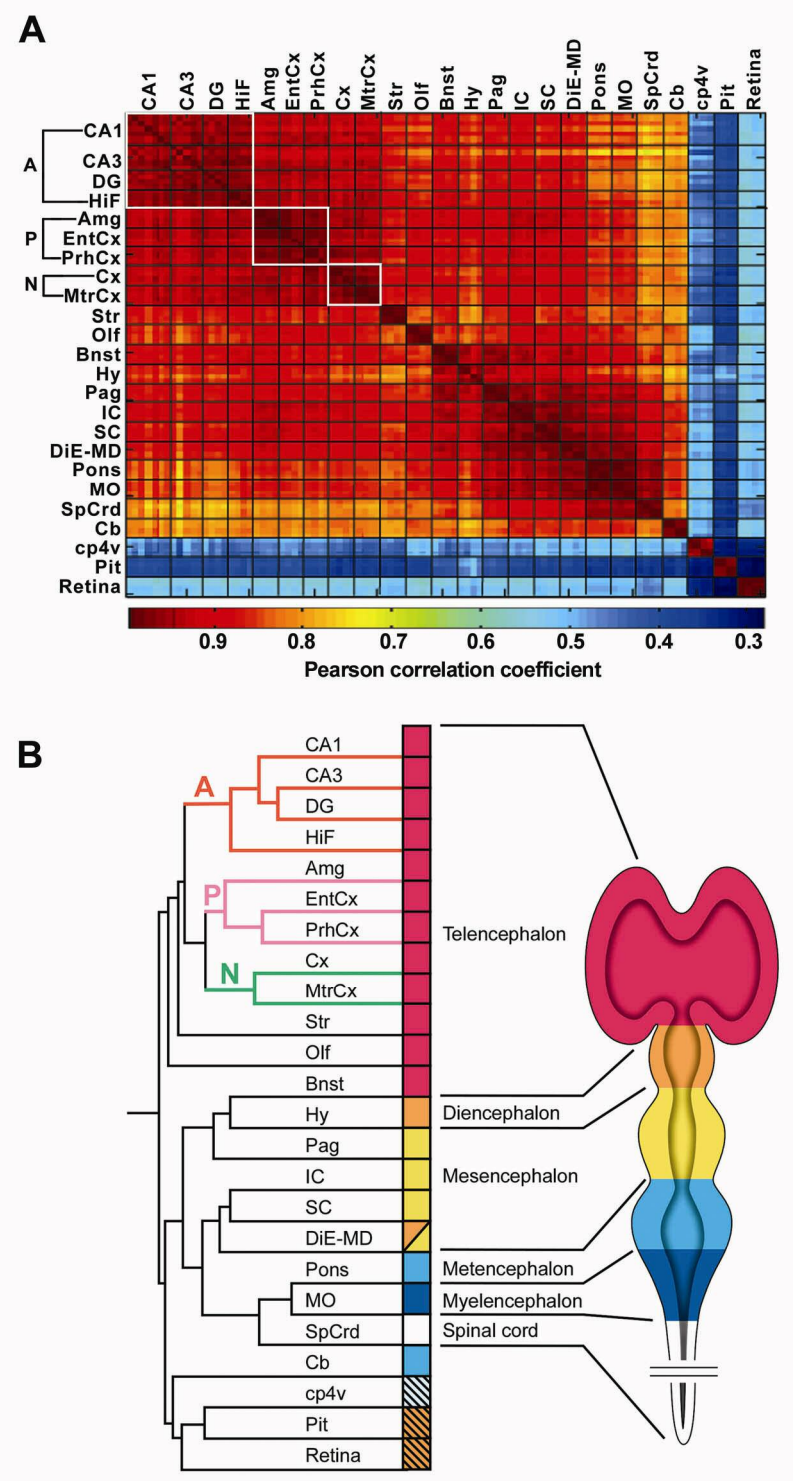
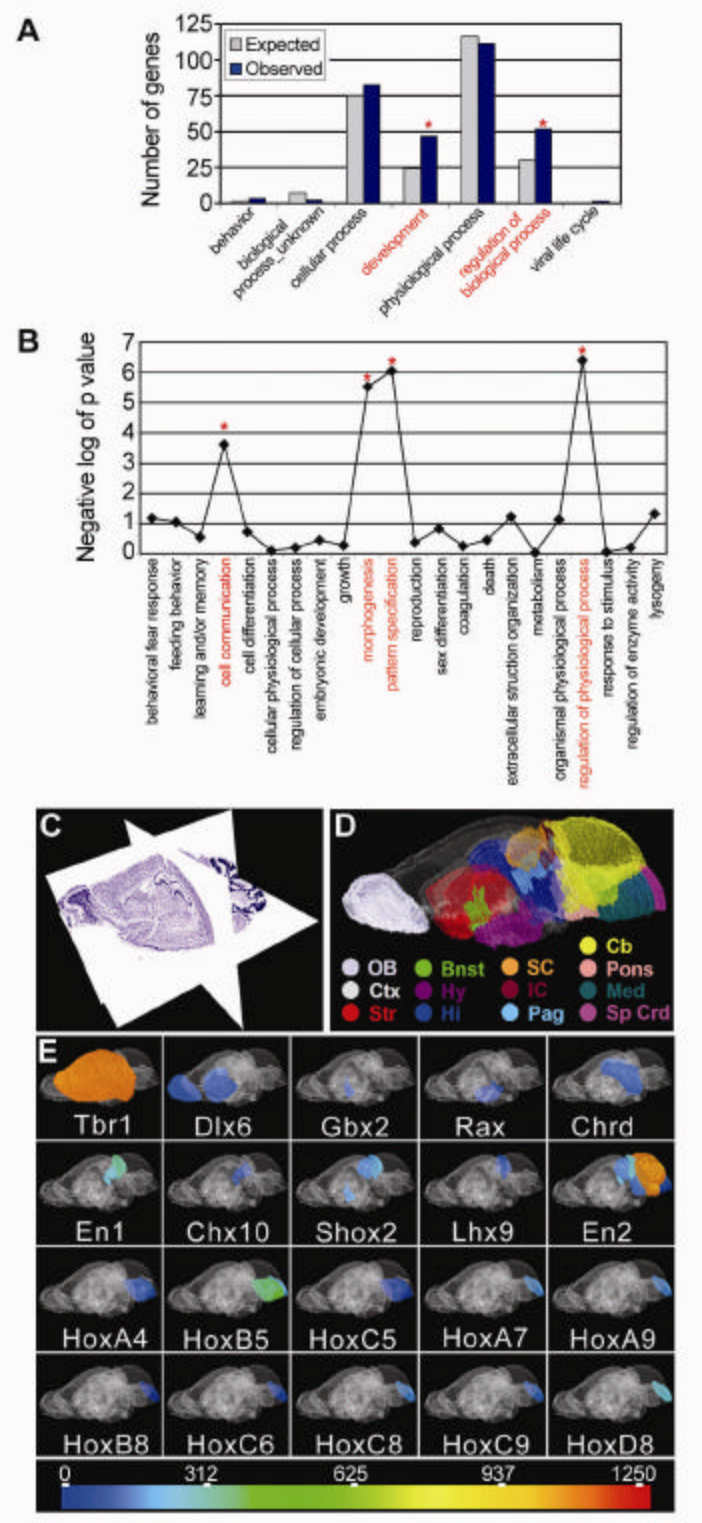


Figure 2. 3-D Brain Atlas. Genes with region-specific expression patterns function in development, pattern specification and morphogenesis. (*A*) The abscissa indicates the functional categories from the Gene Ontology (GO) Tree Machine program. Within the GO Biological Process category, only 'development' and 'regulation of biological process' showed significant over-representation (*$p < 0.01$). The ordinate indicates the number of genes observed in each category compared to the number of genes expected by chance. The significantly over-represented categories are noted by *. (*B*) The GO subcategories in 'development' from (*A*) which are significantly over-represented in the set of genes with region-specific expression patterns. The GO categories are noted along the abscissa; the negative logarithm (base 10) of the p-value is given along the ordinate. Functional categories significantly over-represented are noted by *. (*C*) Reference brain atlas displayed in the three orthogonal planes. This Nissl-stained C57BL/6N mouse brain atlas comprises 462 coronal sections at 30 μm thickness, digitized at a resolution of 1.3 μm/pixel. The sagittal and horizontal planes are "virtual" sections dynamically constructed from the coronal sections. (*D*) Three-dimensional atlas of brain regions. Specific brain regions along the rostrocaudal neuraxis are color coded. (*E*) The expression levels of the homeobox and other embryonic patterning genes expressed in the adult mouse brain are shown for each region. A complete list of these embryonic patterning genes, along with a visual display of their expression on the virtual mouse brain atlas can be found at www.neurome.com/review.

Figure 2. 3-D Brain Atlas



References

Altman J, Bayer SA. (1986). The development of the rat hypothalamus. *Adv Anat Embryol Cell Biol*. 100:1-178.

Barlow C and Lockhart DJ. (2002). DNA arrays and neurobiology--what's new and what's next? *Curr. Opin. Neurobiol.* 12:554-561.

Bayer SA. (1987). Neurogenetic and morphogenetic heterogeneity in the bed nucleus of the stria terminalis. *J Comp Neurol*. 265:47-64.

Bonaventure P, Guo H, Tian B, Liu X, Bittner A, Roland B, Salunga R, Ma XJ, Kamme F, Meurers B, *et al.* (2002). Nuclei and subnuclei gene expression profiling in mammalian brain. *Brain Res.* 943:38-47.

Bowden DM and Martin RF. (1995). NeuroNames Brain Hierarchy. *Neuroimage* 2:63-83.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29:365-371.

Brown VM, Ossadtchi A, Khan AH, Yee S, Lacan G, Melega WP, Cherry SR, Leahy RM, Smith DJ. (2002). Multiplex three-dimensional brain gene expression mapping in a mouse model of Parkinson's disease. *Genome Res*. 12:868-884.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci. USA* 100:13030-13035.

Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, Schambra UB, Nowak NJ, Joyner A, Leblanc G, Hatten ME. *et al.* (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425:917-925.

Holland LZ and Holland ND. (1999). Chordate origins of the vertebrate central nervous system *Curr. Opin. Neurobiol.* 9:596-602.

Jarvis ED, Gunturkun O, Bruce L, Csillag A, Karten H, Kuenzel W, Medina L, Paxinos G, Perkel DJ, Shimizu T, *et al.* (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nat Rev Neurosci.* 6:151-159.

Kamme F, Salunga R, Yu J, Tran DT, Zhu J, Luo L, Bittner A, Guo HQ, Miller N, Wan J, *et al.* (2003). Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J. Neurosci.* 23:3607-3615.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, et al. (2004). Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*. 14:1462-1473.

Lein ES, Zhao X, Gage FH. (2004). Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. *J. Neurosci.* 24:3879-3889.

Lockhart DJ and Barlow C. (2001). Expressing what's on your mind: DNA arrays and the brain. *Nat. Rev. Neurosci.* 2:63-68.

Lumsden A and Krumlauf R. (1996). Patterning the vertebrate neuraxis. *Science* 274:1109-1115.

Pasini A and Wilkinson DG. (2002). Stabilizing the regionalisation of the developing vertebrate central nervous system. *BioEssays* 24:427-438.

Paxinos G and Franklin KBJ. (2001). *The Mouse Brain in Stereotaxic Coordinates*. Acad. Press, San Diego, California, USA.

Puelles L and Rubenstein JL. (2003). Forebrain gene expression domains and the evolving prosomeric model. *Trends Neurosci.* 26:469-476.

Redwine JM, Kosofsky B, Jacobs RE, Games D, Reilly JF, Morrison JH, Young WG, Bloom FE. (2003). Dentate gyrus volume is reduced before onset of plaque formation in PDAPP mice: a magnetic resonance microscopy and stereologic analysis. *Proc. Natl. Acad. Sci. USA* 100:1381-1386.

Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl. Acad. Sci. USA* 97:11038-11043

Shimamura K, Hartigan DJ, Martinez S, Puelles L, Rubenstein JL. (1995). Longitudinal organization of the anterior neural plate and neural tube. *Dev.* 121:3923-3933.

Spemann H and Mangold H. (1924). Über Induktion von Embryonanlagen durch Implantation artfremder Organisatoren. *Roux' Arch. f Entw. mech*. 100:599-638.

Swanson LW. (2000). What is the brain? *Trends Neurosci.* 23:519-527.

Wingate RJ. (2001). The rhombic lip and early cerebellar development. *Curr. Opin. Neurobiol.* 11:82-88.

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol.* 15:1359-1367.

Spemann H. (1918). Über die Determination der ersten Organanlagen des Amphibienembryonen. *Zool. Jahr. Supp.* 15:1-48.

Zapala MA, Lockhart DJ, Pankratz DG, Garcia AJ, Barlow C, Lockhart DJ. (2002). Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biol.* 3:software0001.1-0001.9.

Zhang B, Schmoyer D, Kirov S, Snoddy J. (2004). GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5:16.

Zhao X, Lein ES, He A, Smith SC, Aston C, Gage FH. (2001). Transcriptional profiling reveals strict boundaries between hippocampal subregions. *J. Comp. Neurol.* 441:187-196.

Zirlinger M, Kreiman G, Anderson DJ. (2001). Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. *Proc. Natl. Acad. Sci. USA* 98:5270-5275.

Zirlinger M and Anderson DJ. (2003). Molecular dissection of the amygdala and its relevance to autism. *Genes Brain Behav.* 2:282-294.

CHAPTER 3

DNA variation and brain-region specific expression profiles show different relationships between inbred mouse strains: implications for eQTL mapping studies

Abstract

Expression quantitative trait locus (eQTL) mapping is used to find loci responsible for the transcriptional activity of a particular gene. In recent eQTL studies, expression profiles were derived from either homogenized whole brain or collections of large brain regions. However, the brain is a very heterogeneous organ and expression profiles of different brain regions vary significantly. Because of the importance and potential power of eQTL studies to identify regulatory networks, we analyzed gene expression patterns in different brain regions from multiple inbred mouse strains and investigated the implications for the design and analysis of eQTL studies. Gene expression profiles of five brain regions in six inbred mouse strains were studied. Few genes showed a significant strain-specific expression pattern, while a large number of genes showed brain region-specific patterns. We constructed phylogenetic trees based on the expression relationships between the strains and compared them to a DNA-level relationship tree. The trees based on the expression of strain-specific genes were constant across brain regions and mirrored DNA-level variation. However, the trees based on region-specific genes showed a different set of strain relationships depending on the brain region. An eQTL analysis showed that the majority of strain-specific genes (56%) have *cis*-acting eQTLs. We noticed an enrichment of *cis*-acting regulators among strain-specific genes, while brain region-specific genes seem mainly regulated by *trans*-acting elements. Furthermore, our results suggest that many regulatory networks are highly brain region-specific and indicate the importance of conducting eQTL mapping studies using data from brain

regions or tissues that are physiologically and phenotypically relevant to the trait of interest.

Introduction

Recent genome sequencing efforts have catalogued DNA-level variation between different species, strains, and individuals. In addition, gene expression profiling data indicate that there is considerable variation in expression patterns between strains of inbred mice and individual humans, and several recent articles have studied some of the underlying regulatory mechanisms responsible for this variation (Schadt et al. 2003; Monks et al. 2004; Morely et al. 2004; Chesler et al. 2005; Hubner et al. 2005). The expression studies are based on mapping of so-called "expression quantitative trait loci" (eQTL) in which gene expression profiles are treated as quantitative traits and genome-wide association and linkage mapping are performed to localize regulatory elements that affect the expression of the corresponding differentially expressed genes. The underlying logic is that if a regulatory element coincides with the known location of the differentially expressed gene, it most likely represents a *cis*-acting regulatory element, whereas a regulatory element identified at a different location most likely represents a *trans*-acting regulatory element. However, the relationship between DNA sequence differences and gene expression levels on a genomic scale, and how these two types of variation influence the activities of genes across different tissues has not been studied in detail. We believe that inbred mouse strains offer an excellent model to study the relationship

between DNA-level variation and variation in gene expression patterns as the genealogy and DNA-level variation across different strains are well known. We investigated if inbred strains that are closely related have gene expression profiles that on average resemble each other more than strains that are distantly related. In addition, we were interested in localizing regulatory elements of genes with either strain- or brain region-specific expression patterns by eQTL analyses.

Results

We considered how global DNA-level variation correlates with gene expression pattern variation across five brain regions in six inbred mouse strains. The genealogy of these strains is well known (Beck et al. 2000), and single nucleotide polymorphism (SNP) data are publicly available (Pletcher et al. 2004). We constructed a DNA-level phylogenetic tree based on genetic similarity across 12,473 SNPs (Cervino et al. 2005) (**figure 1a**). The derived relationships correlate well with the known genealogies of the strains and previously published DNA variation-based relationships (Atchley and Fitch 1993; Witmer et al. 2003). We carefully dissected five different brain regions (bed nucleus of the stria terminalis (bnst), hippocampus, hypothalamus, periaqueductal gray (pag), and pituitary gland) from six commonly used inbred mouse strains (129S6/SvEvTac, A/J, C3H/HeJ, C57BL/6J, DBA/2J, and FVB/NJ). Replicate gene expression patterns were measured using the Affymetrix mouse genome 430 2.0 arrays, which contain 45,037 probe sets and cover a significant portion of the mouse transcriptome. Next, we performed a multiple regression

formulation of an analysis of variance using the different mouse strains and brain regions, as well as their interactions, as the independent variables and gene expression signal as the dependent variable to identify genes that exhibited either strain- or region-specific effects. We chose to use a regression model due to the fact that we had an imbalance (61 observations) in our design (Neter et al. 1996). A total of 2,235 probe sets (5.0 %) showed a significant strain-specific effect ($p<0.01$; and the strain effect was more significant than the brain region effect; false discovery rate q-value<0.004). The q-values were obtained using the "smoother method" of Storey and Tibshirani (Storey and Tibshirani 2003). However, even using the more conservative Benjamini and Hochberg (Benjamini and Hochberg 1995) method produces q-values of 0.02 for p-values < 0.01. Somewhat surprisingly, 19,813 probe sets (44.0 %) showed a brain-region-specific expression pattern ($p<0.01$; and the region-specific effect was more significant than the strain effect; q-value<0.001). In addition to the regression formulation that accounted for an unbalanced sample design, a simple 2-way ANOVA in which the outlying unbalanced sample (least correlated) was removed was used in order to determine the number of probe sets that showed a significant strain and brain region interaction. This analysis showed virtually identical results compared to the regression formulation in terms of F-statistics (the F-statistics and p-values of the regression formulation and 2-way ANOVA are available for all probe sets in Supplementary Table 1). The number of probe sets that showed a significant brain region and strain interaction ($p<0.01$; q-value = 0.01) in the 2-way ANOVA model was 7415. These data indicate that while there are significant gene expression

differences between different inbred strains, a large proportion of genes show region-specific expression patterns and interactions between strain and brain region suggesting multiple region-specific regulatory mechanisms controlling gene expression.

In order to determine to what extent the DNA sequence variation correlates to the gene expression level variation in different brain regions, we constructed phylogenetic trees of strain relatedness using either strain-specific or region-specific genes identified by the regression model (**figure 1**). We averaged the (scaled) gene expression signals for the replicate samples for each gene and calculated a Pearson correlation coefficient for the signal intensities between all possible strain combinations for each brain region. We then transformed these correlation coefficients into distances to construct phylogenetic trees (**figure 1**). The tree based on the expression levels of the strain-specific genes (**figure 1c**) has branches that show strain relationships that parallel those based on the SNPs (**figure 1a**). Within each strain, brain-region relationships follow the molecular architecture of the brain (Zapala et al. 2005) shown in **figure 1b**. Likewise, the tree based on the region-specific genes (**figure 1d**) has branches that show individual brain regions clustering according to the molecular architecture. However, the strain relatedness within each brain region branch varies and shows a different set of strain relationships depending on the brain region. Because both the strain-specific and region-specific genes cluster the brain regions according to the known molecular architecture of the brain (Zapala et al. 2005), it is not likely that the observed clustering patterns are due to random noise.

To test if these correlations between the gene expression-based trees and the SNP tree are significant, we broke down the expression trees by brain region, and used Mantel's matrix correspondence test. We compared the strain-specific gene expression trees and the region-specific gene expression trees with the SNP tree for each brain region separately. By using the strain-specific genes, there was a significant correlation between the SNP tree and each of the strain-specific expression trees (bnst R=0.727, p=0.008; hippocampus R=0.680, p=0.002; hypothalamus R=0.529, p=0.008; pag R=0.715, p=0.004; pituitary R=0.512, p=0.023). By contrast, there was no statistically significant correlation between the SNP tree and any of the region-specific expression trees (bnst R=0.466, p=0.180; hippocampus R=0.476, p=0.195; hypothalamus R=0.370, p=0.169; pag R=-0.072, p=0.524; pituitary R=0.271, p=0.135). The strain-specific gene trees were more similar to the SNP tree than the region-specific gene trees (paired t-test p=0.006). When the strain-specific expression trees where compared with each other, all pairwise comparisons (n=10) were statistically significant ($R>0.48$, $p<0.024$). When the region-specific expression trees where compared with each other, only two comparisons out of ten were statistically significant (bnst vs. pit R=0.406, p=0.04 and hippocampus vs. hypothalamus R=0.620, p=0.025), consistent with our proposition that the strain-specific expression trees resemble the SNP tree and each other, and that the region-specific expression trees do not correlate with each other, DNA-level variation or known genealogy. In other words, the known genetic differences (i.e. SNPs between strains) have a low and

insignificant correlation to brain region-specific differences while the strain-specific differences show a high and significant correlation to genetic differences.

These data suggest that because the relatedness of the strains based on strain-specific genes correlate with the DNA-level variation and known genealogy, the expression of strain-specific genes (that comprise only about 5 % of all genes on the array) is mostly regulated by *cis*-acting regulatory elements. DNA variations in a *cis*-regulatory element are likely to mainly affect the transcription of a single gene close to that regulatory element, and more dramatic gene expression differences between strains are associated with *cis*-acting eQTLs (E.E.S., unpublished). Therefore, a phylogenetic tree based on SNPs and a tree based on genes with *cis*-acting regulators should be similar. To assess this hypothesis we performed an eQTL analysis on the gene expression data from the six inbred strains. Indeed, 48 % of the strain-specific probe sets with SNP markers within 4 MB had significant *cis*-acting eQTLs ($p \leq 0.001$) (1,015 out of 2,115 probe sets; a subset of the original 2,235 strain-specific probe sets that had SNP markers located within 4 MB), while only 10 % of the region-specific probe sets showed significant *cis*-acting eQTLs (1,940 of 18,868 region-specific genes with markers within 4 MB). Strain-specific SNPs within a probe set could cause differential hybridization and affect expression results leading to spurious associations and an artificial enrichment of strain-specific *cis*-acting eQTLs. In order to control for strain-specific SNPs that could affect hybridization, we used an algorithm developed in our laboratory that takes advantage of the fact that Affymetrix GeneChips use a series of oligonucleotides that span up to hundreds of bases of a

given gene to detect potential sequence variations between the strains (J.A. Greenhall, M.A.Z., N.J.S., C.B. and D.J.L., manuscript in preparation, see materials and methods). These oligonucleotides (called probes) yield distinct patterns of intensity for each gene. The probe pairs are sensitive enough that appropriately positioned single base differences between the probe pair and the detected RNA can significantly change the signal intensity, and thus produce different patterns between slightly different sequences (Ronald et al. 2005). We compared the underlying patterns of signal intensity between the strains to identify probe sets that may harbor sequence differences. Using a Bonferroni corrected p-value of $p < 0.01$ (calculated from a two-tailed Student's *t*-test (unpaired, equal variance)), 144 out of the 1015 strain-specific probe sets with significant *cis*-acting eQTLs were predicted to harbor sequence differences within the probe set that may affect hybridization. 167 out of the 1940 region-specific probe sets with significant *cis*-acting eQTLs were predicted to harbor sequence differences. When we ignore all probe sets that are predicted to harbor strain-specific sequence differences that could adversely influence hybridization, 56 % of the strain-specific probe sets with SNP markers within 4 MB had significant *cis*-acting eQTLs ($p \leq 0.001$; 901 out of 1611 probe sets), while only 10 % of the region-specific probe sets showed significant *cis*-acting eQTLs (1,773 of 17,422 probe sets). Using a less conservative p-value threshold for the polymorphism detection algorithm did not change the relative enrichment of cis eQTLs among strain specific genes (see Supplementary Table 2).

A caveat of the eQTL analysis is that the limited number of strains leads to a high type I error rate. However, the likelihood that significant false positive eQTLs will be located within 4 MB of the gene of interest, rather than anywhere in the genome, is greatly reduced. Moreover, our eQTL analysis should not be thought of as a traditional eQTL mapping study since it was not focused on the effect of an individual gene or marker, but rather on overall genomic trends or the trends of large groups of genes. For a detailed discussion concerning the determination of the false positive rate, see materials and methods. Our regression model analysis showed that a large proportion of genes that are expressed in the brain are brain-region-specific, and the derived relationships of the strains differed depending on the brain region, suggesting mainly *trans*-acting regulators for these genes, at least in these brain regions. While the eQTL analysis showed a larger number of potentially *trans*-acting eQTLs among the brain region-specific genes (n=3023; compared to 1358 *trans*-acting eQTLs among the strain-specific genes), it is difficult to demonstrate this trend definitively with the small number of strains analyzed.

Our results demonstrate a large number of brain region-specific genes suggesting that many regulatory networks are highly brain region-specific. Certain genes have extremely complicated expression patterns whose variation is dependent on both strain and brain region effects. For example, the relative expression levels for two genes that show significant strain and brain region variation, *preproenkephalin* (*Penk*) and *forkhead box P1* (*Foxp1*) are shown in **figure 2** in a virtual 3D brain atlas. Both genes show interesting strain and region-specific expression patterns. In the

hippocampus and hypothalamus, the expression level of *Penk* is higher in the 129S6/SvEvTac strain compared to the A/J strain. However, in the bnst and in the pag, the expression level of *Penk* is higher in the A/J strain than in the 129S6/SvEvTac strain. Similarly, the expression level of *Foxp1* is higher in the 129S6/SvEvTac hippocampus than in the A/J hippocampus, but in all other regions studied the *Foxp1* expression level is higher in the A/J animals compared to the 129S6/SvEvTac animals.

Discussion

We have shown that the extent of global DNA sequence variation does not directly determine the extent of gene expression variation between inbred mouse strains. Furthermore, the strains that are genetically and genealogically most closely related sometimes have significantly different expression patterns. Interestingly, we observed that the expression of the strain-specific genes seem to be driven mainly by *cis*-acting regulatory elements while the brain region-specific genes are mainly regulated by *trans*-acting regulators. It has been shown that *trans*-acting regulators affect expression levels of multiple genes (Brem et al. 2002), and that both *cis*- and *trans*-acting loci regulate variation in the expression levels of genes, although most act in *trans* (Morley et al. 2005). The heritability estimates for gene expression regulation are relatively low (median value 0.34) (Monks et al. 2004), at least based on expression data from cell lines. Therefore, it is likely that the expression of the majority of genes is influenced by environmental or non-genetic factors including epigenetic mechanisms, such as DNA methylation and histone acetylation.

The large differences in gene expression patterns across the strains depending on the brain region indicate that it is essential to conduct eQTL mapping using data from brain regions that are physiologically and phenotypically relevant for the disease or trait being investigated. Our results show that it is important to dissect a sufficiently small, reasonably homogeneous anatomical region for gene expression profiling studies to avoid "dilution" of strain- and region-specific effects. If several brain regions are combined, the observed gene expression profiles will be a weighted average of the expression profiles of the individual regions. If a gene is expressed at measurable levels in multiple regions, there will be a decrease in the sensitivity to a change in any one region. If there are opposing gene expression patterns in multiple regions, the measurement from a combined sample could miss important changes or even yield misleading information about the underlying regulatory mechanisms.

By investigating DNA polymorphisms and gene expression profiles of various brain regions in six inbred mouse strains, we noticed an enrichment of *cis*-acting regulators among the strain-specific genes, while the brain region-specific genes seem to be mainly regulated by *trans*-acting elements. In addition, our data suggest that different inbred mouse strains have very different relative amounts of certain transcripts in some brain regions, indicating complex brain region-specific regulatory networks. Our findings shed light on regulatory mechanisms of gene expression in different tissues and strains on a genomic scale, and have important implications for the design and analysis of eQTL mapping studies. In order to identify meaningful

regulatory networks, it is important to obtain gene expression profiles from sufficiently small, anatomically refined tissues.

Materials and Methods

Animals

Seven week old male inbred mice were received from the Jackson Laboratory (A/J, C3H/HeJ, C57BL/6J, DBA/2J, and FVB/NJ) or from the Taconic Farms (129S6/SvEvTac). Animals were singly housed for one week before dissections. All animal procedures were performed according to protocols approved by the Salk Institute for Biological Studies Institutional Animal Care and Use Committee.

Tissue collection and RNA preparation for gene expression analysis

All brain dissections were done between 11.00-17.00 h on a petri dish filled with ice using a dissection microscope. The dissected brain regions for gene expression analysis included hypothalamus, hippocampus, pituitary gland, periaqueductal gray, and bed nucleus of the stria terminalis. Hippocampus samples were directly frozen on dry ice and stored at -80^{o}C. The smaller brain structures were collected in RNA Later buffer (Ambion) and samples from 2-5 animals were pooled and stored at -80^{o}C. At least two independent replicate samples for each strain and brain region using independent animals were dissected. If samples were pooled, at least two independent pools were collected. The extraction of total RNA from the

tissues was performed using the TRIzol reagent (Invitrogen) according to the manufacturer's instructions.

Microarray experiments

Gene expression analysis was done using the mouse genome 430 2.0 arrays (Affymetrix) which contain ~45,000 probe sets. Labeling of samples, hybridization, and scanning were performed as described (Zapala et al. 2005). Two replicate samples from independent animals were prepared for each strain and each tissue (the bnst for C3H/HeJ was performed in triplicate).

Data analysis

Array results were analyzed using several different methods. First, .cel files were generated using Affymetrix software, imported into the TeraGenomics expression database, and then processed within the TeraGenomics analysis system (Information Management Consultants) (Zapala et al. 2005). More detailed information on the statistical methods and the TeraGenomics platform can be found in the Supporting Methods and at the TeraGenomics home page (www.teragenomics.com).

Phylogenetic trees were constructed using the UPGMA option of the MEGA3 software (Kumar et al. 2004). SNP trees were constructed based on the fraction of allele differences across all loci between strains. Several different metrics were tested but the tree using this strategy resulted in a tree that correlated best with the known

genealogy of inbred strains. The SNP genotypes were from the same mouse strains as the expression data; except for the 129 strain. We used genotypes from 129S1/SvImJ and gene expression data from 129S6/SvEvTac substrain. We had genotypes available from four different 129 substrains and all of them clustered into a separate clade close to each other in a phylogenetic tree (Cervino et al. 2005). We selected the 129S1/SvImJ genotypes as this strain is genealogically closest to 129S6/SvEvTac. Therefore, the analysis should not have suffered from using a slightly different, but closely related 129 strain for the two types of analyses.

2-factor regression formulations of an analysis of variance were performed with an in-house software program written in standard FORTRAN for Unix using the gene expression files of each array from the absolute analysis of the TeraGenomics analysis system. The results were refined and sorted in Excel. Only genes which scored as Present in one of the files were included in the analysis. In order to test the statistical significance of strain, region, and locus effects on expression levels, we used 2-factor linear regression models. Note that we had independent replicate observations on five mouse brain regions across six mouse strains for a total of 61 observations on the ~45,000 probe sets represented on the microarray (the bnst for C3H/HeJ was performed in triplicate). Let $y_{i,j,k}$ be the expression value of the *i*th replicate (*i*=1,2,) on the *j*th strain (*j*=1,…,6) for the *k*th brain region (*k*=1,…,5). A linear model for the expression values can be written as (Neter et al. 1996):

$$y_{i,j,k} = b_0 + b_{s(1)}x_{i,j,k}(s1) + b_{s(2)}x_{i,j,k}(s2) + b_{s(3)}x_{i,j,k}(s3) + b_{s(4)}x_{i,j,k}(s4) + b_{s(5)}x_{i,j,k}(s5) + b_{r(1)}x_{i,j,k}(r1) + b_{r(2)}x_{i,j,k}(r2) + b_{r(3)}x_{i,j,k}(r3) + b_{r(4)}x_{i,j,k}(r4) + \left[\sum_{s,r} b_{s,r}(\delta_{s,r})\right] + e_{i,j,k}$$

where b_0 is an intercept term, $b_{s(h)}$ is the regression coefficient associated with the effect of the *h*th strain, $b_{r(g)}$ is the regression coefficient associated with the effect of the *g*th brain region, and $e_{i,j,k}$ is an error term. The $x_{i,j,k}(sh)$ and $x_{i,j,k}(rg)$ are indicator variables set to 1 if *ijk*th observation is from strain *h* and/or region *g*, respectively and 0 otherwise. Note that we test only 5 strain and 4 region terms due to redundancy in adding the 6th strain and 5th region in the model. Tests of significance of the strain and region effects involve the hypothesis that the relevant regression coefficient departs from 0.0. Tests of more global hypotheses of *any* strain and/or region effects can be constructed by fitting reduced models that do not include the strain (or region) terms and comparing these reduced models to the 'full' model described above. These global tests involved 5 and 4 degrees-of-freedom for the strain and region effect tests, respectively. We assessed the significance of the difference between the reduced and full models using permutation tests assuming 99 data permutations (i.e., with lowest possible p-value = 0.01). Data was permuted across brain region and strain to determine accurate p-values for the main effects of brain region and strain. To obtain accurate p-values for the interaction terms, the residuals must be permuted which was not done due to increased computational time and complexity (Edgington 1995). Instead, the F-statistics from the resulting regression

model were used to calculate p-values for the cumulative f-distribution; these p-values were also calculated for the strain and brain region effects and used in the False Discovery Rate calculations to calculate the q-values. Note, for the interaction terms, $\delta_{s,r}$, the summation is over all combinations of individual brain regions and strains, such that the $\delta_{s,r}$ simply reflect the product of relevant strain and brain region 0-1 dummy variables. This formulation of interaction terms in regression models is standard in regression contexts. With our regression model, we could have tested each individual regression coefficient in the model for its deviation from 0.0 and hence been able to draw inferences about which brain regions or strains were most likely to deviate from the others in terms of expression level. However, although we included interaction terms in the full model we chose not to focus on them due to potential overfitting and an insufficient number of observations. In order to properly identify interactions, we utilized a 2-way ANOVA calculated using the "anovan" function in Matlab, where the least correlated unbalanced sample was removed. To test hypotheses about individual locus effects, we replaced the strain terms in the full model with a single locus effect (i.e., regression coefficient) term, b_l, and an indicator variable, $x_{i,j,k}(l)$ set to 1 if observation *i,j,k* has a particular allele at locus *l* and 0 otherwise.

Pearson correlation coefficients were calculated using Excel. The formula used to transform correlations into distances is $\sqrt{[2*(1-R)]}$, where R is the correlation coefficient. Mantel's matrix correspondence test was performed with 999

permutations and calculated using GenAlEx 6 (http://www.anu.edu.au/BoZo/GenAlEx/).

eQTL analysis was performed with an in-house software program written in standard FORTRAN for Unix where an F-statistic from a regression model was used at each marker loci to test for an association. A similar 2-factor regression model was used as in the earlier analysis. Results were sorted and analyzed in a separate in-house C++ program. A marker was considered to be *cis*-acting if it was within 4 MB of the start or end position of the gene of interest. 5 MB and 2 MB windows gave similar results. The genomic start and end position of a gene corresponding to the probe set was determined using the Entrez Gene IDs from the Affymetrix database, Netaffx. Both the probe set positions and the SNP marker positions were aligned to NCBI Build 34 (Supplementary Tables 3 and 4). We note that our analysis of *cis* and *trans* acting eQTLs was simply meant to compliment the single degree-of-freedom similarity matrix-based Mantel tests of the hypothesis that similarity in global gene expression patterns do not necessarily correlate with strain DNA sequence similarity, and hence is not meant to unequivocally or definitively identify variations that influence gene expression. It is in this context that we consider what we would expect to observe for our eQTL analyses if no relationship exists between mouse strain and brain region gene expression and the genetic variations the strains possess throughout the genome. To test the association of each locus to each probe set, we used the regression model described above, using the p-value associated with the hypothesis that the regression coefficient, b_l, was equal to 0 in a one degree-of-freedom t-test (no

permutation tests were pursued). We make some simplifying assumptions in our calculations given the difficulty in accounting for correlations between the expression levels of the genes and the haplotype block patterns encompassing the SNPs we examined across the genome. We note that we tested 8680 loci (ignoring monomorphic and missing SNP information, see attached SNP data in Supplementary Table 3) for 22,048 probe sets in our eQTL analysis for a total of 191,376,640 tests of association. We set a p-value threshold of 0.001 to delineate loci worth considering as harboring *cis* or *trans* acting variations. We would thus expect 191,376 of these tests to produce p-values < 0.001 by chance alone if the expression values were independent of each other as well as the relationships between the strains with respect to regulatory variations in their genomes. We observed 3,225,220 associations with p-values < 0.001 – a value much higher than expectation. For the analysis of *cis*-acting eQTLs we note that we included SNPs within four MB of each gene represented by a probe set as being located near enough to the gene to count as possibly *cis*-acting, and, on average, there were 29 SNPs within four MB of each gene. We would expect 29 SNPs x 22,048 probe sets x 0.001 (p-value cutoff) $\approx$ 640 tests to produce p-values < 0.001 by chance alone. We observed 2955 probe sets with p-values < 0.001 for SNPs within four MB of the physical positions of the probe sets.

Polymorphism prediction

Candidate genes harboring predicted polymorphisms were identified using an algorithm developed by our laboratory (J.A.Greenhall, M.A.Z., N.J.S., C.B. and

D.J.L., manuscript in preparation). Briefly, the algorithm works as follows: first, for the selected probe sets, the individual hybridization intensity values are extracted and the difference between the perfect match and the mismatch (PM-MM) intensities is calculated for each probe pair for each sample, excluding probe sets from samples that do not meet certain pattern quality measures. The PM-MM values for each of the probe sets for each sample are globally scaled (by a factor derived from the standard deviation across the multi-probe pattern obtained in each experiment) to compensate for gene expression differences. Next, the scaled values for each sample group are averaged across the strain, and an average and a standard deviation are calculated for each probe pair in a probe set. The appropriate degrees of freedom are calculated and the two-tailed Student's *t*-test (unpaired, equal variance) is derived for each probe pair for each strain comparison. The algorithm was written in C++ and runs on standard UNIX machines. The algorithm has been previously used and validated to identify sequence variation between inbred mouse strains (Carter et al. 2005) and between human, chimpanzee and rhesus macaque (Caceres et al 2003). The algorithm is in principle similar to two previously published methods (Ronald et al. 2005; Borevitz et al. 2003).

3D visualization of gene expression.

Data containing signal intensity values from gene expression microarray analyses were imported in the NeuroZoom software (Neurome, La Jolla).

Visualization of the signal intensities was performed as described earlier (Zapala et al. 2005).

URLs

Further details on the TeraGenomics microarray analysis tool are available at http://www.teragenomics.com. The Affymetrix Netaffx database can be found at http://www.affymetrix.com/analysis/index.affx.

List of Abbreviations

129	129S6/SvEvTac
A	A/J
ANOVA	analysis of variance
bnst	bed nucleus of the stria terminalis
eQTL	expression quantitative trait locus
Foxp1	forkhead box P1
hi	hippocampus
hy	hypothalamus
MB	mega base pair
pag	periaqueductal gray
Penk	preproenkephalin
SNP	single nucleotide polymorphism

Acknowledgements

We thank Information Management Consultants (Reston, VA) for donation of the Teradata data warehouse, design and programming of the TeraGenomics database; Teradata/NCR (Rancho Bernardo, CA) for early support of the project; Barbara Stoveken for help with brain dissections; Floyd Bloom, John Reilly and Warren Young for discussions concerning 3D imaging of brain gene expression; Rick Tennant for help with array hybridizations; and Todd Carter for his insight. We also thank the members of the Barlow laboratory for discussions and technical assistance. This work was supported by the grant MH062344-03 from the National Institute of Mental Health to C.B. and D.J.L, NS039601-04 from the National Institute of Neurological Disorders and Stroke to C.B., and grants from the Academy of Finland to I.H. This chapter appears as a reformatted version of the following published material:

Hovatta I*, **Zapala MA***, Broide RS, Schadt EE, Libiger O, Schork NJ, Lockhart DJ, Barlow C. DNA variation and brain-region specific expression profiles show different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biology*. 2007. 8:R25.

Figure 3. Relationship of inbred mouse strains

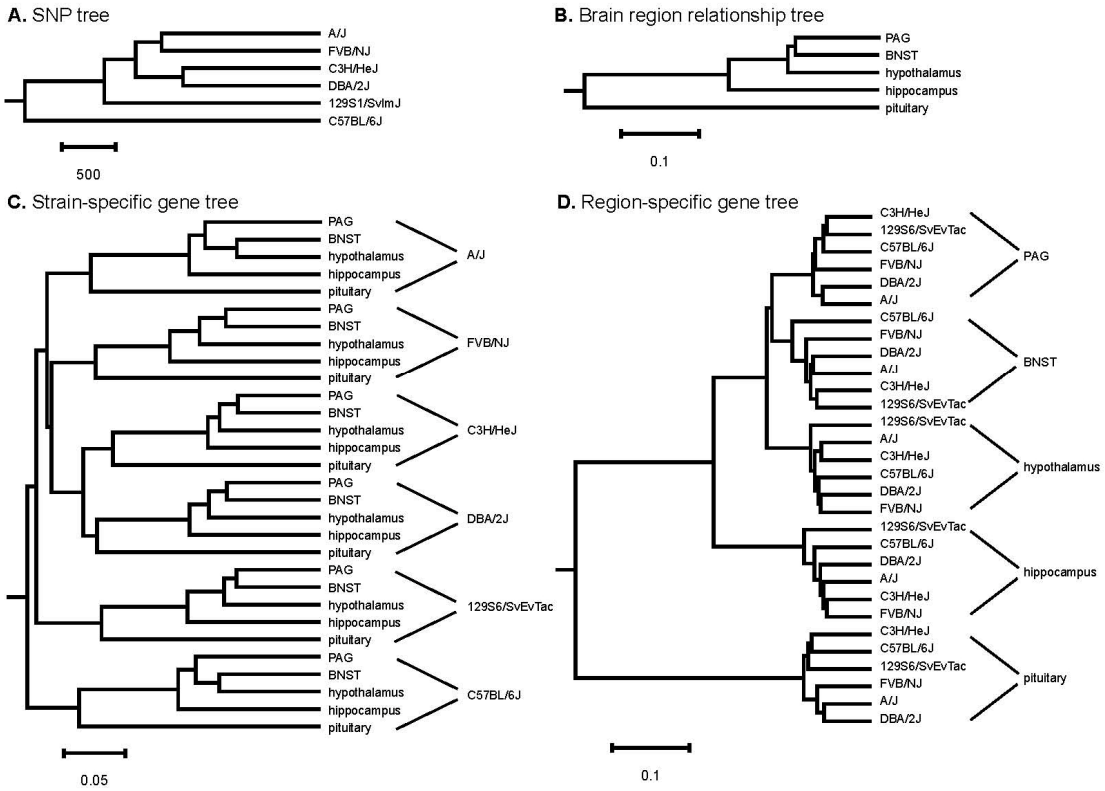
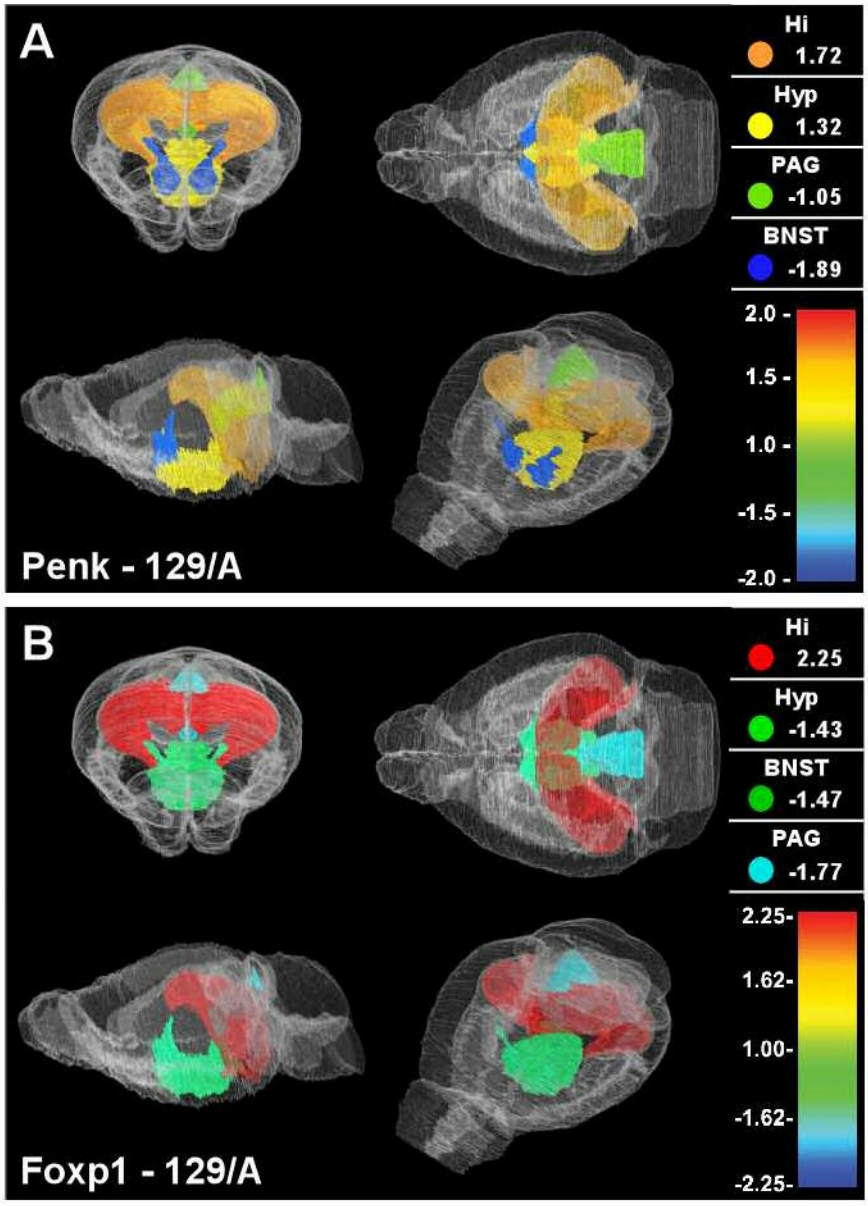


Figure 3. Relationships of inbred mouse strains. A. A phylogenetic tree based on the fraction of allelic differences across 12,473 loci between inbred mouse strains. **B.** A phylogenetic tree based on the gene expression differences between brain regions averaged over six inbred mouse strains used in this study. **C.** A phylogenetic tree based on the gene expression relationship of 2,235 strain-specific genes. **D.** A phylogenetic tree based on the gene expression relationship of 19,813 brain region-specific genes. Scale bars show the number of allelic differences (**A**) or the distance based on gene expression (**B**, **C**, and **D**).

Figure 4. Brain gene expression levels. *Preproenkephalin* (*Penk*) and *forkhead box P1* (*Foxp1*). The signal intensities of two genes, *Penk* and *Foxp1* were imported into the NeuroZoom software tool to visualize the three dimensional gene expression patterns of these genes in the context of brain anatomy. A ratio of the signal intensities of **A.** *Penk* and **B**. *Foxp1* between 129S6/SvEvTac (129) and A/J (A) strains is shown in hippocampus (Hi), hypothalamus (Hyp), periaqueductal gray (Pag), and bed nucleus of the stria terminalis (Bnst). The expression fold change values are shown in the upper right corner of each panel for each brain region separately together with color coding that matches the color of each brain region in the 3D mouse brain atlas, shown from four different angles. Note that the gene expression level of *Penk* in Hi and Hyp is higher in the 129 strain than in the A strain, but in Pag and Bnst it is higher in the A strain compared to the 129 strain. Similarly, the expression level of *Foxp1* in Hi is higher in the 129 strain than in the A strain, while in the Hyp, Bnst, and Pag the expression level is higher in the A strain than in the 129 strain.

Figure 4. Brain Gene Expression Levels



References

Atchley WR, Fitch. (1993). Genetic affinities of inbred mouse strains of uncertain origin. *Mol Biol Evol* 10:1150-1169.

Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF, Fisher EM. (2000). Genealogies of mouse inbred strains. *Nat Genet* 24:23-25.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc* 57:289-300.

Brem RB, Yvert G, Clinton R, Kruglyak L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752-755.

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* 13:513-523.

Carter TA, Greenhall JA, Yoshida S, Fuchs S, Helton R, Swaroop A, Lockhart DJ, Barlow C. (2005). Mechanisms of aging in senescence-accelerated mice. *Genome Biol* 6:R48.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A* 100:13030-13035.

Cervino AC, Li G, Edwards S, Zhu J, Laurie C, Tokiwa G, Lum PY, Wang S, Castellini LW, Lusis AJ et al. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. *Genomics* 86:505-517.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 37:233-242.

Edgington ES. (1995). *Randomization tests*. New York: Marcel Dekker.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* 37:243-253.

Kumar S, Tamura K, Nei M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5:150-163.

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* 75:1094-1105.

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* 430:743-747.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. (1996). *Applied Linear Statistical Models*, 4th edition. Irwin, Chicago.

Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW, Lagler E, Korstanje R, Wang X, Nusskern D, Bogue MA et al. (2004). Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol* 2:e393.

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L. (2005). Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res* 15:284-291.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297-302.

Storey JD, Tibshirani R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100:9440-9445.

Witmer PD, Doheny KF, Adams MK, Boehm CD, Dizon JS, Goldstein JL, Templeton TM, Wheaton AM, Dong PN, Pugh EW et al. (2003). The development of a highly informative mouse Simple Sequence Length Polymorphism (SSLP) marker set and construction of a mouse family tree using parsimony analysis. *Genome Res* 13:485-491.

Zapala MA, Hovatta I, Ellison JA, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS et al. (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A* 102:10357-10362.

CHAPTER 4

Detecting Genetic Variation in Microarray Expression Data

Abstract

The use of high-density oligonucleotide arrays to measure the expression levels of thousands of genes in parallel has become commonplace. To take further advantage of the growing body of data, we developed a method, termed "GeSNP," to mine the detailed hybridization patterns in oligonucleotide array expression data for evidence of genetic variation. To demonstrate the performance of the algorithm, the hybridization patterns in data obtained previously from SAMP8/Ta, SAMP10//Ta and SAMR1TA inbred mice and from humans and chimpanzees were analyzed. Genes with consistent strain-specific and species-specific hybridization pattern differences were identified, and approximately 90% of the candidate genes were independently confirmed to harbor sequence differences. Importantly, the quality of gene expression data was also improved by masking the probes of regions with putative sequence differences between species and strains. To illustrate the application to human disease groups, data from an inflammatory bowel disease study were analyzed. GeSNP identified sequence differences in candidate genes previously discovered in independent association and linkage studies and uncovered many promising new candidates. This approach enables the opportunistic extraction of genetic variation information from new or pre-existing gene expression data obtained with high-density oligonucleotide arrays.

Introduction

High-density oligonucleotide arrays are used routinely to measure quantitative levels of gene expression and to screen thousands of genes for expression differences (Lamb et al. 2006; Lipshutz et al. 1999; Lockhart and Winzeler 2000; Shi et al. 2006). Unlike cDNA microarrays that typically use a single spotted PCR product, Affymetrix oligonucleotide arrays are designed with multiple, different, sequence-specific DNA probes for each gene (Lockhart et al. 1996). The quantitative, multiple-probe hybridization patterns for each gene are reproducible, and the specific patterns depend on the sequence of the DNA or RNA molecules that bind (Chee et al. 1996; Fodor et al. 1993; Fodor et al. 1991; Hacia et al. 1996; Pease et al. 1994; Wodicka et al. 1997). Several papers have highlighted the ability to mine gene expression data from high-density oligonucleotide arrays to find probes that behave unusually within a probe set (Li and Wong 2001) or to find genes with variant splice forms (Hu et al. 2001).

Although gene expression arrays were not designed to detect sequence differences, we reasoned that the underlying multi-probe hybridization patterns could be retrospectively analyzed, probe by probe, to identify possible sequence differences between strains, species or individuals. In addition, the identification and removal or "masking" of probe pairs that target regions with sequence differences should produce more accurate expression results in comparisons between distinct genetic populations (Cáceres et al. 2003; Karaman et al. 2003; Khaitovich et al. 2004; Nagpal et al. 2004). Moreover, global genetic variation screening of this type, along with gene expression profiling, provides a valuable complement to other methods, such as direct candidate

sequencing, SNP genotyping and QTL analysis, for the identification of genes responsible for important phenotypes (Geschwind 2000; Grupe et al. 2001).

Here, we demonstrate the performance of a user-friendly web-based program, "GeSNP" (available at http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi) that can be used to identify potential sequence variation from gene expression datasets. This algorithm was used previously to identify sequence differences between three rare strains of inbred mice (Carter et al. 2005) and to improve the reliability of gene expression data by masking probe pairs that cover regions with sequence differences between humans and chimpanzees (Cáceres et al. 2003). Recently, this algorithm was also applied in an expression QTL (eQTL) study to exclude spurious *cis*-acting eQTLs due to probe-specific hybridization differences (Hovatta et al. 2007). Sequence variation that affects hybridization to the array probes and leads to false associations has been identified as a major problem in eQTL investigations (Peirce et al. 2006; Radcliffe et al. 2006). Sequencing of specific candidate genes has been used to minimize these false associations, but this approach is not practical on a genome scale (Hubner et al. 2005). In addition to our studies, similar techniques have also been used to identify sequence differences in prokaryotes and lower eukaryotes (Albert et al. 2005; Borevitz et al. 2003; Gresham et al. 2006; Hazen et al. 2005; Ronald et al. 2005; Winzeler et al. 1998) and between primate species (Khaitovich et al. 2004). However, a fully-functioning program for use with data from many different species and different Affymetrix arrays has not been made available.

Results

Development of the algorithm

On an Affymetrix oligonucleotide array, each gene is represented by a probe set, which is comprised of approximately eleven to twenty different oligonucleotide probe pairs that are designed to hybridize to specific regions of a gene. Each probe pair consists of a matched set of two 25-base oligonucleotide probes, a perfect match (PM) for the gene of interest and a mismatch (MM) containing a single nucleotide substitution in the middle of the probe (position 13). The MM serves as a measure of nonspecific background binding and noise. The collection of signal intensities for all the probes in a probe set is defined as the hybridization pattern. The GeSNP algorithm compares these detailed hybridization patterns across the oligonucleotide probe pairs for a gene, after normalizing for expression level differences, in order to find probe pairs that show consistent, statistically significant differences between two sets of samples. The algorithm works as follows (see Figure 1): first, the individual hybridization intensity values are extracted from the cell-by-cell intensity (CEL) file. The difference between the perfect match and the mismatch (PM-MM) intensities is calculated for each probe pair for each CEL file. In order to minimize false predictions of sequence differences due to inadequate hybridization signal, a probe set from a particular CEL file is excluded if fewer than 65% of the PM-MM values for the probe set are positive, indicating that the gene was not likely expressed at a detectable level. After eliminating data from samples that do not fulfill this criterion, the PM-MM values for all of the probe sets for each sample are globally scaled to compensate

for gene expression differences. The scaling factor is calculated by dividing an arbitrary target value of 200 by the standard deviation of the PM-MM values for a probe set while ignoring the largest and smallest PM-MM values. Next, the scaled values for each sample group are averaged, and an average and a variance are calculated for each probe pair in a probe set. To further reduce false positives, only probe sets for which at least four files in both sample groups have exceeded the pattern quality threshold are analyzed.

To identify statistically significant pattern differences, the Student's *t*-test using the separate variance formula was employed. The *t*-value for each probe pair (PP) was calculated as follows:

$$\frac{\left|\text{Average Scaled Difference}(\text{PP}_a, \text{group}_1) - \text{Average Scaled Difference}(\text{PP}_a, \text{group}_2)\right|}{\sqrt{\left(\frac{\text{Variance}(\text{PP}_a, \text{group}_1)}{n_1} + \frac{\text{Variance}(\text{PP}_a, \text{group}_2)}{n_2}\right)}}$$

where n_1 is the number of files in group 1 and n_2 is the number of files in group 2. Empirically, the *p*-value of the Student's *t*-test and *p*-values generated by permutation testing did not perform as well as the *t*-value in identifying confirmed sequence differences. In order to choose an appropriate *t*-value threshold for each comparison, a false positive comparison was created in which the two compared groups were equally distributed between two false positive groups, accounting for any subgroup bias. Probe pairs that are statistically significant in this randomized comparison represent the potential number of false positives at a particular threshold for the number of files compared (see Methods). Previous studies have referred to the probe pairs containing

putative sequence differences as Single Feature Polymorphisms (SFPs) (Winzeler et al. 1998), and for consistency we adopt this notation.

Identification of sequence differences between three rare mouse strains

We have been studying the aging of the mouse brain using three strains of Senescence Accelerated Mice (SAM) developed in Japan as models of accelerated aging, SAMP8/Ta (SAMP8), SAMP10//Ta (SAMP10) and the control, aging-resistant strain, SAMR1TA (SAMR1) (Takeda 1999). Initially, large-scale sequence or SNP information was not available for these strains, as is still the case for numerous other laboratory strains or crosses. To identify sequence differences between these three strains, MG-U74Av2 Affymetrix arrays hybridized to 5 hippocampal and 4 retinal samples of each mouse strain (Carter et al. 2005) were analyzed using the GeSNP algorithm. Figure 2 shows the number of SFPs identified at increasing *t*-value thresholds, and these numbers reflect the known genetic divergence between the three strains from microsatellite markers (Xia et al. 1999) and SNP genotyping (Cervino et al. 2005). The performance of the GeSNP algorithm was determined by sequencing 24 genes from two cortex samples of each strain and calculating the number of true positives, false positives, true negatives and false negatives at various *t*-value thresholds. Table 1 shows the algorithm performance for the SAMP8 vs. SAMR1 comparison. The true positive rate, also known as the positive predictive value, is the percentage of SFPs that are indeed true positives, and the detection rate, also known as the sensitivity, describes the percentage of probe pairs covering sequence differences

that are identified as SFPs by the algorithm. When results are analyzed by probe pair, a *t*-value of 6 yields an 89% true positive rate and a 75% detection rate. However, most false negatives and false positives at the probe pair level are contained within a probe set harboring true positives, and hence the performance at the probe set level yields a 100% true positive rate and a 100% detection rate for almost all *t*-values shown. A small number of false positives is acceptable in this type of screen as the purpose is to identify a manageable number of candidates for follow-up studies with conventional DNA sequencing. For the SAMP8 vs. SAMR1 and SAMP10 vs. SAMR1 probe sets containing one or more SFPs with a *t*-value $\geq$ 5, see Supplementary Tables 1-2.

Identification of sequence differences between species

In order to further validate the algorithm and extend its applicability to inter-species comparisons, gene expression data from ten humans (*Homo sapiens*) and seven chimpanzees (*Pan troglodytes*) obtained with human HG-U95Av2 Affymetrix arrays were analyzed (Enard et al. 2002; Cáceres et al. 2003). A total of 28 arrays hybridized to human samples and 21 to chimpanzee samples were compared, and the sequences of 24 human and chimpanzee genes (represented by 37 different probe sets) were examined. All but one of the genes were selected for sequencing because they appeared to be differentially expressed between primate brains in the initial analysis of the array data (Cáceres et al. 2003). The performance of the algorithm in the human-chimpanzee comparison is shown in Table 2 (see Supplementary Table 3 for probe

sets containing SFPs with a *t*-value ≥ 5). At a *t*-value threshold of 6, 16 of 19 probe sets (84% true positive rate) or 42 of 59 probe pairs (71% true positive rate) identified as SFPs were independently confirmed to contain sequence differences (Table 2). Similarly, 414 of 431 probe pairs (96% specificity) covering identical sequence between species did not contain SFPs. In addition, taking into account the quality of the current genome sequences, comparable results were obtained based on a genome-wide analysis of the available human and chimpanzee genome sequence (see Supplementary Table 7).

Compared to the analyses of closely related mouse strains, the detection rates are lower in this inter-species comparison. Approximately 50% of false negatives are due to the sequence difference lying within the first five or last five nucleotides of the probe sequence. When comparing populations that are not isogenic, the greater genetic variation within a group can make these more subtle hybridization differences at the ends of the probes difficult to detect. Increasing the number of individuals in each sample group can improve the detection rate and increase the resolution of the data, decreasing false positives and increasing true positives.

Comparison of the GeSNP algorithm to other methods

The performance of the GeSNP algorithm was compared to the algorithm of Ronald et al. (2005) which uses a different approach for background subtraction and normalization and was previously used to identify sequence differences in gene expression data between yeast strains. James Ronald and Leonid Kruglyak kindly provided us with their C++ program implementing both the normalization procedure

of Irizarry et al. (2003) and PerfectMatch of Zhang et al. (2003), which is a positional-dependent-nearest-neighbor model that uses probe target nucleotide sequence and position to determine background binding. Comparing *t*-values from each algorithm, the GeSNP algorithm performed better for a test data set comparing SAMP10 and SAMR1 at all thresholds (see Table 3). At *t*-value thresholds of 6 and 7, the improvement in performance was significant using a chi-square test ($p = 0.0027$ and $p = 0.024$, respectively). These results show that in identifying confirmed sequence differences, the GeSNP algorithm is more robust and accurate than similar computational methods. Although not discussed here in detail, in addition to single nucleotide differences, the GeSNP algorithm was also able to detect insertions/deletions and splice variants.

Improving the quality of array-based gene expression data

Sequence differences in gene regions covered by the oligonucleotide probes can affect the quantitative measurement of expression levels. This effect is especially important when mRNA from one species is interrogated with arrays designed for a different species. The ability to identify probes that cover regions with sequence differences and eliminate them from the analysis is essential for accurate gene expression quantification (Cáceres et al. 2003; Karaman et al. 2003; Khaitovich et al. 2004). For example, in the comparison of the human and chimpanzee gene expression data, we identified 246 probe sets that initially appeared to be expressed at different levels between human and chimpanzee brains

(Cáceres et al. 2003). However, once the signals for the probe pairs predicted to have sequence differences were ignored or "masked," 53 of these probe sets were no longer scored as being differentially expressed. One of the genes, *CTNNA1*, was independently determined by quantitative RT-PCR to be expressed at the same level in humans and chimpanzees (Cáceres et al. 2003). Sequence variation also affected probe hybridization and led to false assignments of differential expression in the SAM comparisons. For example, *Fgf1* gene expression levels appeared lower in SAMP10 mice than in SAMR1 mice. However, GeSNP identified sequence differences in four *Fgf1* probe pairs. Sequencing confirmed that nucleotide differences were covered by the *Fgf1* probes and further led to the identification of a functionally important mutation outside of the probe set region (Carter et al. 2005). Quantitative RT-PCR showed no gene expression difference between SAMP10 and SAMR1 mice, and after masking the affected probe pairs, *Fgf1* was no longer considered differentially expressed. Finally, sequence variation that affects probe hybridization was identified in another mouse study comparing C57BL/6J to 129S6/SvEvTac mice for the gene *Kcnab2* (Sandberg et al. 2000). Therefore, although the occurrence of this phenomenon is more frequent in inter-species studies, the potential effects on gene expression measurements due to sequence differences should be routinely investigated.

Identification of disease-causing mutations in human disease groups

In order to test the ability of the GeSNP algorithm to identify sequence differences between human disease populations, human gene expression data from a study on inflammatory bowel disease were analyzed (Burczynski et al. 2006). Data for peripheral blood samples on Affymetrix HG-U133A arrays were obtained from GEO accession number GSE3365 and directly from the authors. The aim of the study was to identify gene expression signatures from peripheral blood mononuclear cells that could discriminate between two common inflammatory bowel diseases, Crohn's disease and ulcerative colitis. Both disorders are thought to result from genetic and environmental factors that lead to an abnormal immune response in the gastrointestinal tract, with Crohn's disease having a larger genetic component than ulcerative colitis (Sartor 2006).

The data of 59 Crohn's disease patients, 26 ulcerative colitis patients and 42 healthy controls were analyzed with the GeSNP algorithm in order to identify potential sequence differences between these groups. Several previously identified Crohn's disease and ulcerative colitis candidate susceptibility genes showed SFPs, including *SLC22A4*, identified in linkage and association studies (Giallourakis et al. 2003; Ma et al. 1999; Mirza et al. 2003; Rioux et al. 2000; Rioux et al. 2001; Sartor 2006; Waller et al. 2006; among others) and showing functional genetic variants (Peltekova et al. 2004), and *TLR4* (Franchimont et al. 2004; Gazouli et al. 2005; Noble et al. 2006) and *IL1RN* (Carter et al. 2001; Tountas et al. 1999), both of which have been associated with Crohn's disease and ulcerative colitis in certain

populations. Thus, the GeSNP analysis was able to identify several well-known candidate genes. In addition, many promising new candidates that could be involved in inflammatory bowel disease pathogenesis were also identified (see Supplementary Tables 4-6), including two interesting but as yet unimplicated genes, *VIL2* and *HMGB1,* and *F2RL1*, for differences between Crohn's disease and ulcerative colitis.

Discussion

The results demonstrate that the GeSNP algorithm can identify sequence differences using array-based gene expression data. The approach is general to several Affymetrix gene expression array types and applicable to the analysis of data obtained in different populations of genetically distinct individuals, including humans. With most array designs, the sequence coverage for each gene is incomplete. Usually 100 to 400 bases of sequence are interrogated for each gene since there are typically eleven to twenty probe pairs per gene, the probes are 25 bases in length, some of the probes are overlapping, and sequence differences that result in mismatches near the probe ends (e.g., the five bases at either end) are not expected to lead to consistently measurable hybridization differences (Chee et al. 1996; Pease et al. 1994). Nonetheless, this approach allowed us to take advantage of previously existing data, obtained initially for other purposes, to search in a broad and unbiased way for genetic differences without the need for any additional experiments. The GeSNP program can

be used not only to identify small sequence differences, such as single-base substitutions, but also larger deletions or insertions and genes with different splice forms (Hu et al. 2001; Li and Wong 2001; Winzeler et al. 1998). We further illustrated the additional information that can be generated with publicly available data files that contain detailed clinical or phenotypic information. Using GeSNP, we identified several well-known inflammatory bowel disease candidate genes and many new, promising candidates that are consistent with the disease pathophysiology. Thus, this analysis method can be used to complement gene expression and other more traditional studies to accelerate the identification of genes that may mediate important diseases and phenotypes.

In addition to the identification of genetic variants, the analysis methods described here may find their most immediate application in improving array performance and enabling arrays designed for one strain or species to be used more broadly. We have used this technique successfully in the past to improve the quality of gene expression data by masking probes that cover regions with potential sequence differences in both mouse (Carter et al. 2005) and human studies (Cáceres et al. 2003). Identifying sequence variation that may influence hybridization patterns and lead to incorrect results is even more critical in eQTL analysis. There is growing interest in using eQTL studies to discover loci genetically associated with gene expression differences and to determine transcriptional regulatory networks (Bystrykh et al. 2005; Chesler et al. 2005; Schadt et al. 2003). However, SNPs within a probe region that affect expression results might be in linkage disequilibrium with a marker SNP and

lead to a false eQTL association with the marker (Peirce et al. 2006). As eQTL studies become even more prominent, methods that minimize false positive associations will be increasingly important.

In addition, as new array designs become more widely used, the GeSNP algorithm could have a much larger impact. For example, Affymetrix recently released the exon arrays to interrogate all putative exons in a genome. The human array contains 1.4 million probe sets with four PM probes per set (5.6 million probes, 140 million nucleotides). Assuming that half of the probe sets pass the pattern quality measure of detectable expression and that only 50% of the covered nucleotides provide information due to probe sequence overlap and lower sensitivity to differences at the probe ends, analysis with the GeSNP algorithm could yield information on approximately 35 million bases of sequence. However, because specific MMs for each probe are not part of the new exon array design, the pattern quality control and background subtraction techniques would need to be modified in order to apply GeSNP to these arrays.

In summary, the GeSNP algorithm allows for the unbiased, opportunistic extraction of sequence variation information from array-based gene expression data. This information can be used to improve the quality of gene expression and eQTL analyses and to identify potential disease-causing genes in human disease populations. The GeSNP source code and a web-based program are available for public implementation.

Methods

Computer software

The algorithm was written in standard ANSI C++ and compiled to run on UNIX. The extensively commented source code is available for download from the Genome Research website and the GeSNP website, http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi. In addition, the GeSNP website hosts a user-friendly web-based tool that allows users to upload their expression data in two pre-defined groups and obtain results online. A user manual and example data are also available at the website. The GeSNP program outputs a text file for each comparison with the following columns: Probeset, Probepair, pspp (probe set with probe pair number appended at the end), N1 (number of files included in group one), Mean1, Var1 (variance of group 1), N2 (number of files included in group two), Mean2, Var2 (variance of group 2) and *t*-value. In the *t*-value column, the value "NaN" means that there were less than two files included in one or both groups and a *t*-value could not be calculated.

False Positive Estimation and Choosing a Threshold

Because the *p*-value of the Student's *t*-test and a permuted *p*-value with 100,000 permutations did not perform as well as the *t*-value alone in correctly identifying sequence differences, we developed a method to obtain an approximate, "predicted" true positive rate in order to determine an appropriate *t*-value threshold.

First, a false positive comparison is generated, where the two groups of interest are equally distributed into two false positive groups, accounting for any subgroup bias in tissue type, race, gender or prominent diseases. Ideally, no differences should be identified between these two randomized groups. The number of PPs exceeding a *t*-value threshold for the false positive comparison yields an estimated number of false positives for the specific files and number of samples being compared. Subtracting the number of estimated false positives at a given *t*-value from the number of putative SFPs, then dividing by the number of SFPs, yields a predicted true positive rate. The larger the number of independent samples for a comparison, the more accurate the results, assuming no subgroup bias is introduced. For studies within a species with homozygous loci, at least four independent samples should be used in each group. For studies within outbred populations, use of at least ten independent samples per group is advisable.

Analysis of SFPs

Result files were filtered in Microsoft Access according to the minimum number of files (N1 and N2 $\geq$ 4) and the *t*-value threshold. In the supplementary tables, the data are organized by probe set to illustrate important summary information. The number of SFPs in a probe set and the largest *t*-value of these SFPs (with at least one positive mean) are shown. A larger number of SFPs within a probe set, a greater *t*-value and/or multiple probe sets representing a single gene provide increased confidence that a true sequence difference exists for that gene. Annotation

files were downloaded from Affymetrix (http://www.affymetrix.com/analysis/index.affx). Additional information on candidate genes was obtained from NCBI's Entrez, OMIM and PubMed.

Since increasing the number of files can improve analytical power, we also compared combined groups. For example, we combined ulcerative colitis and Crohn's disease samples and compared this group to control samples. While all the SFPs in common to both the ulcerative colitis vs. normal and Crohn's disease vs. normal lists appear in the combined comparison, additional probe sets are identified as SFPs. These probe sets may have more subtle hybridization differences (such as, sequence differences near the ends of the probes) that are enhanced with the larger number of files.

Implementation of the Ronald et al. (2005) algorithm

James Ronald and Leonid Kruglyak kindly provided their C++ program (pdse.cpp) implementing some of the normalization procedure of Irizarry et al. (2003) and PerfectMatch of Zhang et al. (2003). We then wrote a program in MatLab to follow the remaining methods outlined in Ronald et al. (2005). Using the output of pdse.cpp for the SAMP10 vs. SAMR1 comparison, the MatLab program divided the observed intensity by the expected intensity (while expected intensity > 100) and then calculated group means, group variances and the *t*-values between groups.

RNA preparation and cDNA synthesis for sequence confirmation

Total RNA was prepared using Trizol Reagent (Gibco/BRL) following the manufacturer's recommended protocol. For SAM strains, RNA was extracted from the cortex of at least two separate mice for each strain. Standard protocols were used for the generation of cDNA from RNA. Primers were designed to amplify the regions defined by the Affymetrix probe set target sequences of the selected genes, which can be downloaded from the Affymetrix Analysis Center web-site. Standard PCR reactions were performed on an Applied Biosystems GeneAmp PCR System 9700, and PCR products were purified using the recommended procedures for the QIAquick PCR purification kit protocol or the QIAquick gel extraction kit protocol (Qiagen). All sequencing was performed by the Salk Institute Sequencing Core. The sequences of human genes were obtained from GenBank. Chimpanzee sequences were described in Cáceres et al. (2003) and were also obtained from GenBank. For the global comparison of the HG_U95Av2 Affymetrix array probes to the human and chimpanzee genomes, we used the sequence assembly versions HG18 (NCBI Build 36.1) and Pantro2 (Build 2, Version 1). Probe sequences were aligned to the genome sequences using MegaBlast and only probes with 100% identity over the 25 nucleotides were selected.

Acknowledgements

We would like to thank Stephen Heinemann of the Salk Institute for support, Charles Abney for web programming assistance, Eva Mitter of IMC for technical assistance, Sebastian Fuchs for sequencing assistance, Jo Del Rio for preliminary analysis, James Thomas for help with the human and chimpanzee genome analysis, and Svante Pääbo, Todd Carter and Michael Burczynski for providing data files. JAG was supported by a generous gift from the Lewin family and the Sprint Corporation, the NIH Neuroplasticity of Aging Training Grant (5 T32 AG00216) and the National Defense Science and Engineering Graduate Fellowship. MC was supported by an EMBO Long-Term Fellowship, a Salk Institute Innovation Grant and the Ramón y Cajal Program (Ministerio de Educación y Ciencia, Spain). Additional funding was provided by the DOD grant DAMD17-99-1-9561 and the Frederick B. Rentschler Developmental Chair to CB. This chapter appears as a reformatted version of the following published material:

Greenhall JA*, **Zapala MA***, Caceres M, Libiger O, Barlow C, Schork NJ, Lockhart DJ. Detecting genetic variation in microarray expression data. *Genome Research*. 2007. In Press.

Figure 5. Detection of sequence using expression data

a) Key steps in the GeSNP algorithm are described (left panels in boxes), and corresponding graphical illustrations of the SAMP10 data for MG-U74Av2 array probe set 98333_at, representing the gene ribosomal protein S18 (*Rps18),* are shown on the right. Step 1 of the method is to extract data for a specific probe set from the CEL file. In step 2, the hybridization intensity difference between the perfect match and mismatch probe (PM-MM) for each probe pair (PP) is calculated. These values are then evaluated for inclusion in subsequent analyses as determined by passing pattern quality measures for detectable expression. The unscaled hybridization intensity values for *Rps18* are shown for all nine samples of the SAMP10 strain, where the PP number is indicated on the x-axis ranging from 1 to 16, and the PM-MM value is shown on the y-axis. Next (step 3), the intensity patterns for each sample are individually scaled to a common value. The scaled PP differences are then averaged (step 4) to generate a single value and standard deviation for each PP.

b) For the *Rps18* probe set, the same analysis was performed for the nine SAMR1 samples, all of which passed the pattern quality measures for detectable expression. The average hybridization patterns with standard deviations obtained for SAMP10 (red line and squares) and SAMR1 (blue line and triangles) mice are shown. Using a *t*-value threshold of 6, the algorithm identified two PPs harboring putative sequence differences (black asterisks). Consistent with the hybridization pattern differences, DNA sequencing showed that each of these PPs indeed covered a region with a single base pair difference between the two strains.

c) The average hybridization signals with standard deviations are shown for the 96498_at probe set for the gene, disrupted meiotic cDNA 1 homolog (*Dmc1h*), using the six files that passed the pattern quality measures for SAMP10 (red line and squares) and five files for SAMR1 (blue line and triangles). DNA sequencing identified no sequence differences between strains, consistent with the nearly identical, overlapping hybridization patterns (largest *t*-value of 2.4).

Figure 5. Detection of sequence using expression data

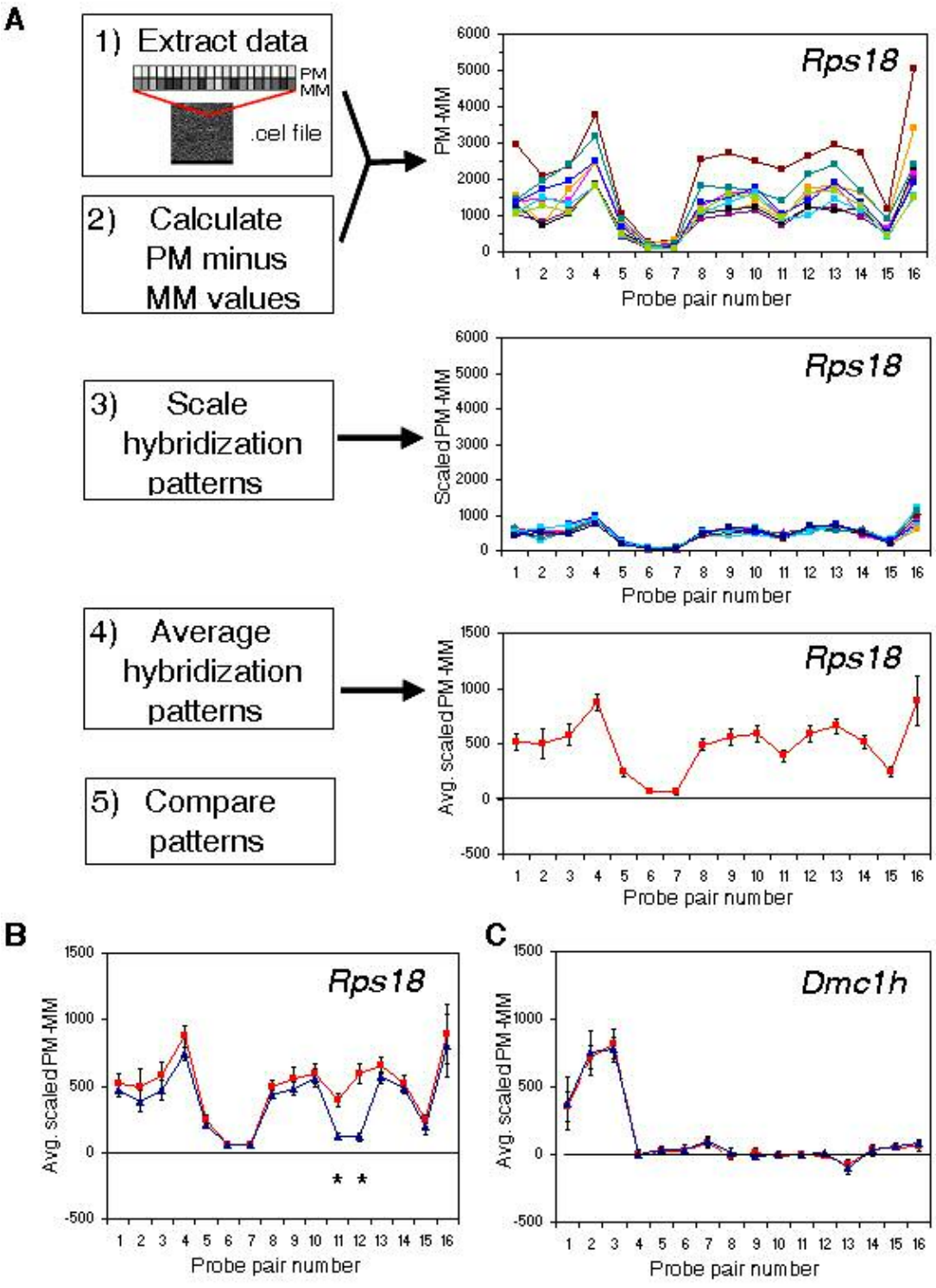


Figure 6. Sequence differences between SAM strains

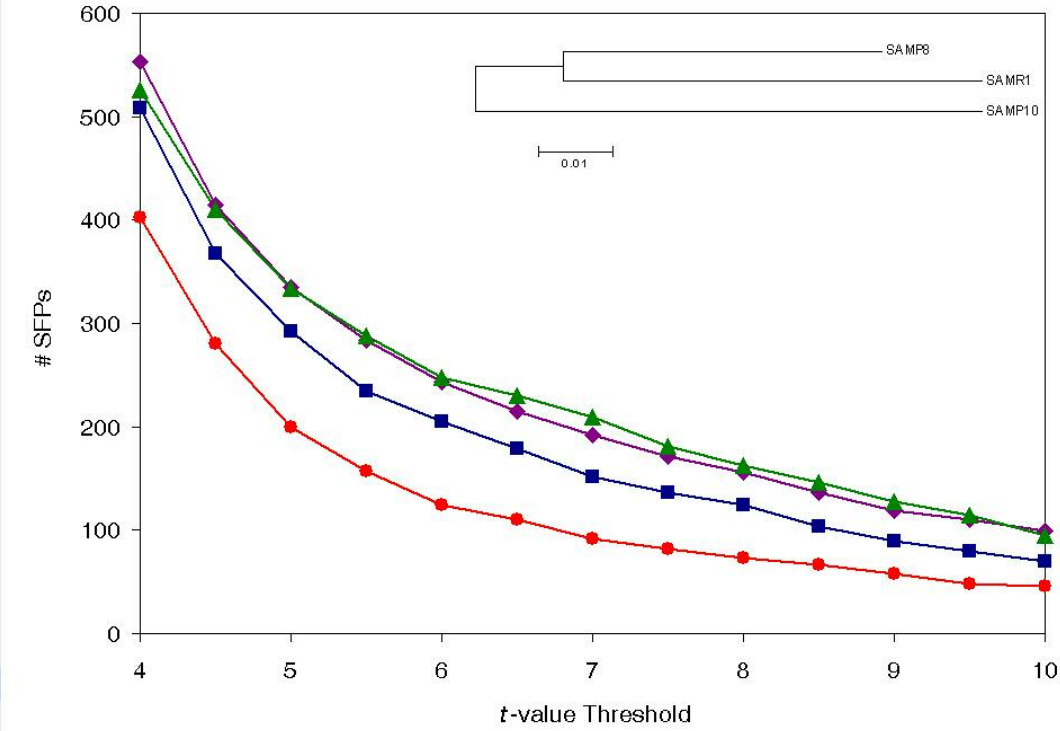


Figure 6. Sequence differences between SAM strains

GeSNP comparisons using the three SAM strains are shown. The y-axis represents the number of SFPs occurring at a given *t*-value, and the x-axis shows several *t*-value thresholds. The SAMP8 to SAMR1 comparison is represented as blue squares. The SAMP10 to SAMR1 comparison is shown as purple diamonds. The SAMP10 to SAMP8 comparison is illustrated as green triangles, and the comparison of SAMP8 and SAMP10 to SAMR1, uncovering SAMP8 and SAMP10 common, shared differences, is shown as red circles. The divergence in the number of SFPs is consistent with the phylogenetic distance between strains that has been found by microsatellite studies (Xia et al. 1999) and SNP genotyping (Cervino et al. 2005). All the non-monomorphic SNPs (1907) of Cervino et al. (2005) were used to generate the phylogenetic tree shown inside the graph using the neighbor-joining option of the MEGA3 software (Kumar et al. 2004).

Table 1. GeSNP performance, SAMP8 vs. SAMR1 mouse strains

TP, true positive; FN, false negative; TN, true negative; FP, false positive; TPR, true positive rate (the positive predictive value), [TP/(TP+FP)]; DR, detection rate (the sensitivity), [TP/(TP+FN)].

		$t \geq 5$	$t \geq 6$	$t \geq 7$
By Probe pair	**TP**	24	24	21
	FN	8	8	11
	TN	365	367	369
	FP	5	3	1
	TPR	0.83	0.89	0.95
	DR	0.75	0.75	0.66
By Probe set	**TP**	7	7	6
	FN	0	0	1
	TN	19	19	19
	FP	0	0	0
	TPR	1.00	1.00	1.00
	DR	1.00	1.00	1.00

Table 2. GeSNP performance, human vs. chimpanzee

TP, true positive; FN, false negative; TN, true negative; FP, false positive; TPR, true positive rate (the positive predictive value), [TP/(TP+FP)]; DR, detection rate (the sensitivity), [TP/(TP+FN)].

		$t \geq 5$	$t \geq 6$	$t \geq 7$
By Probe pair	**TP**	54	42	36
	FN	44	56	62
	TN	395	414	422
	FP	36	17	9
	TPR	0.60	0.71	0.80
	DR	0.55	0.43	0.37
By Probe set	**TP**	19	16	16
	FN	4	8	9
	TN	8	10	10
	FP	6	3	2
	TPR	0.76	0.84	0.89
	DR	0.83	0.67	0.64

Table 3. SAMP10 vs. SAMR1 performance, GeSNP vs. Ronald et al. (2005)

TP, true positive; FN, false negative; TN, true negative; FP, false positive; TPR, true positive rate (the positive predictive value), [TP/(TP+FP)]; DR, detection rate (the sensitivity), [TP/(TP+FN)]; *, significantly different using a chi-square test.

		GeSNP			Ronald et al. (2005)		
		$t \geq 5$	$t \geq 6$	$t \geq 7$	$t \geq 5$	$t \geq 6$*	$t \geq 7$*
By Probe pair	**TP**	21	21	20	10	8	8
	FN	8	8	9	11	13	13
	TN	377	387	388	314	315	320
	FP	12	2	1	12	11	6
	TPR	0.75	0.91	0.95	0.45	0.42	0.57
	DR	0.72	0.72	0.69	0.48	0.38	0.38
By Probe set	**TP**	7	7	7	6	5	5
	FN	1	1	1	2	2	2
	TN	18	19	19	18	18	18
	FP	1	0	0	1	2	2
	TPR	0.88	1.00	1.00	0.86	0.71	0.71
	DR	0.88	0.88	0.88	0.75	0.71	0.71

References

Albert TJ, Dailidiene D, Dailide G, Norton JE, Kalia A, Richmond TA, Molla M, Singh J, Green RD, Berg DE. (2005). Mutation discovery in bacterial genomes: metronidazole resistance in Helicobacter pylori. *Nat. Methods* 2: 951-953.

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res*. 13: 513-523.

Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, Casciotti L, Maganti V, Reddy PS, Strahs A, Immermann F, et al. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J. Mol. Diagn.* 8: 51-61.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, et al. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet*. 37: 225-232.

Cáceres M, Laucher J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci.*. 100: 13030-13035.

Carter MJ, Di Giovine FS, Jones S, Mee J, Camp NJ, Lobo AJ, Duff GW. (2001). Association of the interleukin 1 receptor antagonist gene with ulcerative colitis in Northern European Caucasians. *Gut* 48: 461-467.

Carter TA, Greenhall JA, Yoshida S, Fuchs S, Helton R, Swaroop A, Lockhart DJ, Barlow C. (2005). Mechanisms of aging in senescence-accelerated mice. *Genome Biol*. 6: R48.

Cervino AC, Li G, Edwards S, Zhu J, Laurie C, Tokiwa G, Lum PY, Wang S, Castellini LW, Lusis AJ, et al. (2005). Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. *Genomics* 86: 505-517.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. (1996). Accessing genetic information with high-density DNA arrays. *Science* 274: 610-614.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, et al. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 37: 233-242.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, et al. (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340-343.

Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. (1993). Multiplexed biochemical assays with biological chips. *Nature* 364: 555-556.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* 251: 767-773.

Franchimont, D, Vermeire S, El Housni H, Pierik M, Van Steen K, Gustot T, Quertinmont E, Abramowicz M, Van Gossum A, Deviere J, et al. (2004). Deficient host-bacteria interactions in inflammatory bowel disease? The toll-like receptor (TLR)-4 Asp299gly polymorphism is associated with Crohn's disease and ulcerative colitis. *Gut* 53: 987-992.

Gazouli M, Mantzaris G, Kotsinas A, Zacharatos P, Papalambros E, Archimandritis A, Ikonomopoulos J, Gorgoulis VG. (2005). Association between polymorphisms in the Toll-like receptor 4, CD14, and CARD15/NOD2 and inflammatory bowel disease in the Greek population. *World J. Gastroenterol*. 11: 681-685.

Geschwind DH. (2000). Mice, microarrays, and the genetic diversity of the brain. *Proc. Natl. Acad. Sci*. 97: 10676-10678.

Giallourakis C, Stoll M, Miller K, Hampe J, Lander ES, Daly MJ, Schreiber S, Rioux JD. (2003). IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *Am. J. Hum. Genet*. 73: 205-211.

Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham MJ, Botstein D, Kruglyak, L. (2006). Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* 311: 1932-1936.

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. (2001). In silico mapping of complex disease-related traits in mice. *Science* 292: 1915-1918.

Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. (1996). Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat. Genet*. 14: 441-447.

Hazen SP, Borevitz JO, Harmon FG, Pruneda-Paz JL, Schultz TF, Yanovsky MJ, Liljegren SJ, Ecker JR, Kay SA. (2005). Rapid array mapping of circadian clock and developmental mutations in Arabidopsis. *Plant Physiol*. 138: 990-997.

Hovatta I, Zapala MA, Broide RS, Schadt EE, Libiger O, Schork NJ, Lockhart DJ, Barlow C. (2007). DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biol*. 8: R25.

Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y. (2001). Predicting splice variant from DNA chip expression data. *Genome Res*. 11: 1237-1245.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet*. 37: 243-253.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 31: e15.

Karaman MW, Houck ML, Chemnick LG, Nagpal S, Chawannakul D, Sudano D, Pike BL, Ho VV, Ryder OA, Hacia JG. (2003). Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. *Genome Res*. 13: 1619-1630.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, et al. (2004). Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*. 14: 1462-1473.

Kumar S, Tamura K, Nei M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform*. 5: 150-163.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313: 1929-1935.

Li C, Wong WH. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci*. 98: 31-36.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. (1999). High density synthetic oligonucleotide arrays. *Nat. Genet*. 21: 20-24.

Lockhart DJ, Dong H, Byrne MC, Follettie KT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech*. 14: 1675-1680.

Lockhart DJ, Winzeler EA. (2000). Genomics, gene expression and DNA arrays. *Nature* 405: 827-836.

Ma Y, Ohmen JD, Li Z, Bentley LG, McElree C, Pressman S, Targan SR, Fischel-Ghodsian N, Rotter JI, Yang H. (1999). A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflamm. Bowel Dis*. 5: 271-278.

Mirza MM, Fisher SA, King K, Cuthbert AP, Hampe J, Sanderson J, Mansfield J, Donaldson P, Macpherson AJ, Forbes A, et al. (2003). Genetic evidence for interaction of the 5q31 cytokine locus and the CARD15 gene in Crohn disease. *Am. J. Hum. Genet*. 72: 1018-1022.

Nagpal S, Karaman MW, Timmerman MM, Ho VV, Pike BL, Hacia JG. (2004). Improving the sensitivity and specificity of gene expression analysis in highly related organisms through the use of electronic masks. *Nucleic Acids Res*. 32: e51.

Noble C, Nimmo E, Gaya D, Russell RK, Satsangi J. (2006). Novel susceptibility genes in inflammatory bowel disease. *World J. Gatroenterol*. 12: 1991-1999.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci*. 91: 5022-5026.

Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, Belknap JK, Rosen GD, Goodwin S, Sutter TR, Williams RW, et al. (2006). How replicable are mRNA expression QTL? *Mamm. Genome* 17: 643-656.

Peltekova VD, Wintle RF, Rubin LA, Amos CI, Huang Q, Gu X, Newman B, Van Oene M, Cescon D, Greenberg G, et al. (2004). Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat. Genet*. 36: 471-475.

Radcliffe RA, Lee MJ, Williams RW. (2006). Prediction of cis-QTLs in a pair of inbred mouse strains with the use of expression and haplotype data from public databases. *Mamm. Genome*. 17: 629-642.

Rioux, JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, et al. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet*. 29: 223-228.

Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, et al. (2000). Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am. J. Hum. Genet*. 66: 1863-1870.

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L. (2005). Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res*. 15: 284-291.

Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl. Acad. Sci*. 97: 11038-11043.

Sartor RB. (2006). Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol*. 3: 390-407.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422: 297-302.

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY, et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 24: 1151-1161.

Takeda T. (1999). Senescence-accelerated mouse (SAM): a biogerontological resource in aging research. *Neurobiol. Aging* 20: 105-110.

Tountas NA, Casini-Raggi V, Yang H, Di Giovine FS, Vecchi M, Kam L, Melani L, Pizarro TT, Rotter JI, Cominelli F. (1999). Functional and ethnic association of allele 2 of the interleukin-1 receptor antagonist gene in ulcerative colitis. *Gastroenterology* 117: 806-813.

Waller S, Tremelling M, Bredin F, Godfrey L, Howson J, Parkes M. (2006). Evidence for association of OCTN genes and IBD5 with ulcerative colitis. *Gut* 55: 809-814.

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, et al. (1998). Direct allelic variation scanning of the yeast genome. *Science* 281: 1194-1197.

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol*. 15: 1359-1367.

Xia C, Higuchi K, Shimizu M, Matsushita T, Kogishi K, Wang J, Chiba T, Festing MF, Hosokawa M. (1999). Genetic typing of the senescence-accelerated mouse (SAM) strains with microsatellite markers. *Mamm. Genome* 10: 235-238.

Zhang L, Miles MF, Aldape KD. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*. 21: 818-821.

CHAPTER 5

Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables

Abstract

A fundamental step in the analysis of gene expression and other high-dimensional genomic data is the calculation of the similarity or distance between pairs of individual samples in a study. If one has collected *N* total samples and assayed the expression level of *G* genes on those samples, then an *N x N* similarity matrix can be formed which reflects the correlation or similarity of the samples with respect to the expression values over the *G* genes. This matrix can then be examined for patterns via standard data reduction and cluster analysis techniques. We consider an alternative to conventional data reduction and cluster analyses of similarity matrices that is rooted in traditional linear models. This analysis method allows predictor variables collected on the samples to be related to variation in the pair-wise similarity/distance values reflected in the matrix. The proposed multivariate method avoids the need for reducing the dimensions of a similarity matrix, can be used to assess relationships between the genes used to construct the matrix and additional information collected on the samples under study, and can be used to analyze individual genes or groups of genes identified in different ways. The technique can be used with any high-dimensional assay or data types and is ideally suited for testing subsets of genes defined by their participation in a biochemical pathway or other *a priori* grouping. We showcase the methodology using three published gene expression data sets.

Introduction

The introduction of high-throughput technologies, such as DNA microarrays and proteomics platforms, has provided researchers with a set of assays that are unprecedented in their sophistication. These technologies allow researchers to interrogate the expression levels of thousands to tens-of-thousands of genes or proteins simultaneously (Lockhart et al. 1996; Gygi et al. 1999). Although of tremendous importance, the use of these technologies is plagued by the fact that they generate enormous amounts of data, whose significance, both statistically and biologically, can be difficult to fathom in any single experiment (Bassett et al. 1999). In essence, the collection of expression levels on thousands of genes on relatively few individuals or other units of observation, such as cells or cell types, creates enormous potential for false positive results when each gene is analyzed in isolation (Storey and Tibshirani 2003).

Many clever and useful data analysis strategies for the assessment of gene expression and related high-dimensional genomic data have been proposed (Allison et al. 2006). The vast majority of these strategies rely on either some form of data reduction, such as cluster analysis (Eisen et al. 1998), or eigenstructure analysis (Alter et al. 2000; Alter et al. 2003), which raises a number of questions about the appropriateness of the cluster method used, the number of clusters or eigenvalue/eigenvector pairs seen as "optimal," appropriate or statistically significant, as well as the biological meaning of the clusters or eigenvectors that emerge (Bryan 2004). Despite this fact, one common and appropriate strategy exploited by a number

of data analysis approaches – which is in fact a precursor and fundamental construct to many contemporary gene expression analysis methods – involves the construction of a similarity or distance matrix, which reflects the similarity/dissimilarity of each pair of individuals with respect to the gene expression values obtained on them. This strategy was outlined in many of the earliest proposed gene expression analysis methods (Allison et al. 2006 ; Spellman et al. 1998; Alon et al. 1999, Golub et al. 1999), has become a standard tool for gene expression data analysis and visualization tools (Slonim 2002; D'Haeseleer 2005), and is, in fact, even a typical ingredient in cluster and eigenstructure analyses.

We describe a method for testing the relationship between variation in a distance matrix and predictor information collected on the samples whose gene expression levels have been used to construct the matrix. The method provides a formal test of the organization of a similarity or distance matrix as it relates to predictor variable information collected on the individual samples, such as clinical parameters on subjects whose tumors have been evaluated for gene expression or genotype data of different inbred mouse strains assayed for gene expression. As a result, the method is the perfect companion for heat map and tree-based representations of high-dimensional data organized by some feature or *a priori* grouping factor meant to graphically represent and reveal a relationship between the genes used to construct the heat map or tree and these features or groupings. By testing more global hypotheses about the patterns within a similarity or distance matrix, the procedure avoids the need for cluster analysis and is very appropriate for

situations where the number of data points collected is much larger than the number of samples or individuals. We first describe the derivation of the method, and then showcase its application to three publicly available datasets. We also want to emphasize that the procedure can be used to study any number of groups of genes, including single genes or all the genes in a data set, making it very flexible and a method that only adds to existing univariate approaches.

Methods

The Basic Model

In describing the proposed analysis methodologies, we follow the notation in McArdle and Anderson (2001). We do not focus on many of the alternative methodologies for distance-based analyses developed by Krzanowski (2002), Gower and Krzanowski (1999), Legendre and Anderson (1999), and Gower and Legendre (1986), although many of these techniques may have some merit in the analysis of genomic data. Note that we used boldface to indicate matrices or vectors in our notation. Let **Y** be an N x P matrix harboring gene expression values on N subjects for P genes. Let **X** be an N x M matrix harboring information on M predictor or regressor variables whose relationship to the gene expression values is of interest, where the first column contains a column vector whose every element is 1, and reflects an intercept term, as in standard regression contexts. These predictor variables could include the ages of individuals assayed, clinical diagnoses, strain memberships, cell line types, or genotype information. A standard multivariate multiple regression model for this situation would be (Johnson and Wichern 1992; Anderson 2001):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \quad (1)$$

where $\boldsymbol{\beta}$ is an M x P matrix of regression coefficients and ε is an error term, often thought be distributed as a (multivariate) normal vector. The least-squares solution for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, with the matrix of residual errors for the model being

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad (2)$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Unfortunately, If $N<<P$, as is often the case with gene expression and other genomic data types, then this model is problematic. An alternative would consider how the M predictor variables relate to the similarity or dissimilarity of the subjects under study with respect to the P gene expression values as a whole or as a series of unique subsets of the data.

Let **D** be an N x N distance matrix, whose elements, d_{ij}, reflect the distance (or dissimilarity) of subjects i and j with respect to the P gene expression values. For example, d_{ij} could be calculated as the Euclidean distance or as a function of the correlation coefficient (see the section on *Forming the Distance Matrix* below). Let $\mathbf{A} = (a_{ij}) = (-[1/2]d_{ij}^{\,2})$. One can form Gower's centered matrix G from A by calculating:

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \quad (3)$$

where $\mathbf{1}$ is a N-dimensional column vector whose every element is 1 and $\mathbf{I}$ is an N *x* N identity matrix. An appropriate F-statistic for assessing the relationship between the M

predictor variables and variation in the dissimilarities among the *N* subjects with respect to the *P* variables is:

$$F = \frac{tr(\mathbf{HGH})/(M-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-M)} \quad (4)$$

Where **H** is a hat matrix, **G** is Gower's centered matrix, and **I** is the identity matrix, formed as above. *M* is scalar and reflects the number of predictors and *N* is the number of subjects. If *P* = 1 (i.e., a univariate analysis) and the distance matrix is computed through the use of the standard Euclidean distance measure, then *F* in equation (4) is the standard *F*-statistic and possesses the typical properties associated with *F*-statistics in ANOVA contexts. This result is due to the fact that the inner product matrix (*Y'Y*) used in standard univariate analysis of variance and regression contexts contains the same information, in terms of total sums-of-squares, as the outer product matrix (*YY'*) which reflects interpoint squared differences or distances ($tr(Y'Y) = tr(YY')$) (McArdle and Anderson 2001). When different distance measures are used, the properties of *F* are more complicated, suggesting the use of alternative methods for assessing statistical significance (see the section on *Testing Statistical Significance* below).

Forming the Distance Matrix

The formation of the distance matrix is an important step in the use of the proposed procedure. There is a bewildering array of potential distance measures one could use with the proposed method (Webb 2002). The correlation coefficient, *r*, is

often used to assess the similarity between two individuals based on gene expression values (D'Haeseleer 2005). A correlation matrix with elements r_{ij} can be converted to a distance matrix with elements d_{ij} easily enough through a simple transformation:

$$d_{ij} = \sqrt{2(1 - r_{ij})} \tag{5}$$

This transformation leads to a distance matrix with metric properties, although distance measures with non-metric properties can be used in the analysis method described as well (Gower and Krzanowski, 1999). We discuss aspects of the choice of a distance measure in the results section, but more work in this area is needed. One additional aspect of the formation of a distance matrix that deserves attention involves handling missing data. Intuitively, if one has collected thousands of gene expression values to be used to create distance profiles then a few missing observations are not likely to have much of an influence. For example, one could simply not use these genes in the formation of the distance matrix, ignore the missing values only when assessing pair-wise distance for a pair of observations with missing data, or assign individuals with missing data imputed values that are then used to construct distance. However, the delineation of a threshold beyond which the number of missing values creates problems for a distance-based analysis is important and worthy of further research.

Assessing Statistical Significance

The distribution of the *F*-statistic defined in equation 4 is complicated and its derivation for any particular distance matrix is unlikely to generalize to other distance

matrices, especially with small sample sizes. Therefore, one can rely on permutation tests to evaluate the probabilistic significance of an observed *F*-statistic computed from equation 4 (Edgington 1995; Manly 1997; Good 2000). Permutations can either involve permuting the raw data or simultaneously permuting the rows and columns of the **G** matrix, as is done in Mantel's matrix correspondence test (McArdle and Anderson 2001). In addition, if permutation tests are used, the degrees-of-freedom terms in the numerator (*M*-1) and the denominator (*N*-*M*) are not required in the formulation of the statistic presented in equation 4. Finally, given that different predictor variables, or subsets of variables, can be tested for association with variation in a distance matrix, one can pursue step-wise or variable selection procedures with the technique, identical to univariate standard multiple regression analysis (Neter et al. 1985). Beyond a *P*-value, an estimation of the proportion of variation within the matrix that is explained by a particular set of *M* predictor variables can be calculated by dividing $tr(\mathbf{HGH})$ (i.e. the sum of the diagonal elements of a matrix) by $tr(\mathbf{G})$. In our analyses, independent variables are tested both individually and in a forward stepwise manner. The independent variables selected for the model in a stepwise manner are based on the highest cumulative proportion of variance that is explained by the inclusion of an additional variable in the regression model. An *F*-statistic and *P*-value are calculated for the addition of this variable to the model.

Assessing Level Accuracy and Power of the Proposed Hypothesis Test

In order to examine properties of the proposed analysis procedure, a series of studies investigating the level accuracy and power of the proposed test statistics were performed. A more complete investigation of the properties of the test procedure is forthcoming. In order to examine the level accuracy of the test, we simulated 30 samples each measured on 100 variables. The variables were assumed to follow a standard normal distribution; hence, there was no structure to the data. We assumed that the first 15 samples had a different origin than the second 15. We then tested the relationship between this grouping factor (coded as 0 for the first 15 samples and 1 for the second 15 samples) and the distance between the samples using different distance measures using the proposed procedure and 1000 random permutations of the data. We repeated this process 1000 times. Table 1 describes the results and clearly shows that the nominal level of the test matches closely with the simulation results, suggesting that the proposed test procedure is non-biased.

We also considered the power of the proposed test. We simulated data for 30 samples and 100 variables in which 15 samples were assigned to a hypothetical control group (independent variable = 0) and 15 samples were assigned to a hypothetical experimental group (independent variable = 1). Data in the control group were generated as standard normal variables with a mean of 0 and variance 1. Data in the experimental group were generated as standard normal variables with variance = 1 and means that took on values of 0 to 1.5 in increments of 0.001. The power of the proposed permutation-based statistical test was then investigated in these settings. In

this context we also generated different simulated data sets for which 100%, 50%, 25%, 10% or 5% of the variables used in the construction of the distance matrix had means adjusted from 0 (in the appropriate increments) in the experimental group. Figure 1 describes the results. Note that the grey line in the figure represents the power curve obtained based on a t-test with the Bonferroni correction, corrected for 100 multiple comparisons. Figure 1 clearly shows that the proposed procedure can detect "signals" in the data as long as the number of variables contributing to that signal used in the construction of the distance matrix is moderate.

Results

The proposed method was tested on three different published data sets to display its utility. We briefly consider some of the implementation details and properties of the proposed technique, such as the need for evaluating the distance between the observations, and the dependence of the test statistic on subsets of genes among all those used to derive a distance measure. We note that for the following applications we used the correlation coefficient to derive the distance measure, as this measure has been the standard for gene expression data (D'Haeseleer 2005). In addition, we used 1000 permutations to compute *P*-values.

The Embryonic Imprint of the Adult Mouse Brain

The first data set involved gene expression data from multiple brain regions and multiple inbred mouse strains (Zapala et al. 2005). The normalized data can be

downloaded from GEO using record number GSE3594. The authors had three hypotheses about the relationships of the gene expression patterns between the different brain regions in the adult mouse. The gene expression patterns of these brain regions could be related to each other based on adult anatomy, evolutionary relationships, or embryonic origin. The authors performed hierarchical cluster analysis and created a Pearson correlation heat map matrix where they hypothesized that the gene expression patterns of the adult mouse brain bear an imprint based on the adult tissue's embryological origin. The heat map and the hierarchical tree constructed based on the similarity of gene expression patterns across all the genes suggest that adult structures are related to each other based on the classic five vesicle embryonic neural regions (telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon; see supplementary Figure 1). Using the proposed regression analysis procedure we provide statistical evidence that embryonic origin is the most likely hypothesis of the three since it explains the largest proportion of variation in the similarity of the overall gene expression profile of the brain regions (P-value < 0.001 and proportion of variation in pair-wise distances explained by embryological origin = 0.33, adult anatomy = 0.26, and evolutionary relationships = 0.19). The authors also suggested that anterior to posterior (A/P) position along the neural tube could dictate expression patterns in the adult neural structures. The position along the neural tube was tested individually and in combination with embryological origin. Importantly, A/P position added a significant proportion of explained variation in brain region gene

expression similarity over and above embryological origin (*P*-value < 0.001 and proportion of variation explained above embryological origin = 0.032).

The Aging Human Brain

The second data set examined gene expression patterns in the human frontal cortex among individuals who died at various ages (26-106 years) (Lu et al. 2004). The normalized data can be downloaded from GEO using record number GSE1572. The authors performed Spearman rank correlations to determine 463 genes that correlated with age (*P*-value < 0.005) and a Pearson correlation-based heat map matrix was then calculated that covered all pair-wise comparisons of individuals. We analyzed the distance matrix based on the pair wise correlations (equation 5) using the proposed regression method to quantify the effect that age and sex may have on the gene expression patterns for the 463 genes found to be correlated with age (age *P*-value < 0.001 and proportion of variation explained = 0.35; sex *P*-value = 0.224). While sex was not a significant predictor of the gene expression patterns in the frontal cortex, age appeared to explain approximately 35% of the variation in the similarity in the gene expression patterns among the individuals based on the age-related genes (see Supplementary Figure 2A). Moreover, the association with age was not only apparent in age-related genes (as identified from Spearman rank correlations), but was also evident in the correlation matrix created using all the genes scored as "Present" in at least one of the samples (age *P*-value < 0.001 and proportion of variation explained = 0.16; sex *P*-value = 0.78) (see Supplementary Figure 2B). Therefore, it appears that

the age effect is pronounced enough to be significant even when including all detectable genes on the array (8507 probe sets).

As emphasized, the proposed regression technique can be applied to each gene in a univariate manner. Univariate analysis can be used to identify a set of genes that, when considered in the construction of a distance matrix, are strongly related to a specific predictor variable. For example, we wanted to identify a set of age associated genes that would explain a larger proportion of variation than the set that Lu et al. identified by using Spearman Rank correlations. Using the matrix regression technique for each individual gene (using the Euclidean distance) we calculated an *F*-statistic which, as noted when Euclidean distances are used with a single variable, is identical to an ANOVA statistic (McArdle and Anderson 2001). The resulting *F*-statistics were then used to rank genes and identify those which demonstrated the largest age association (Figure 2A, 2B). These ranked genes were then serially used to construct matrices tested with our proposed procedure to identify an optimal set of genes that resulted in the largest proportion of variation explained by age effects (Figure 2C, 2D). An optimal set was found to occur with the top 100 ranking genes, as age explained 52% of the variation in dissimilarity within the matrix constructed with these genes, which is much higher than the 35% explained with the age-related genes chosen by Lu et al. (2004). Of the 100 genes identified as an optimal set, 80 were within the Lu et al. set. Even using an equivalent number of genes as Lu et al. used, the proportion of variation explained by our highest ranked 463 genes is 0.42 compared to the Lu et al. proportion of 0.35. Of our 463 highest ranked genes, only 256 are also present in the

Lu et al study. Interestingly, among the 20 genes within the top 100 genes identified by our analysis that were not identified in the Lu et al. study were two genes involved specifically with neurological function, mitogen-activated protein kinase kinase 1 (*MAP2K1*) and amyloid beta (A4) precursor protein-binding, family A, member 1 (*APBA1*). *MAP2K1* is known to control apoptosis signalling specifically in astrocytes (Gomez Del Pulgar et al. 2002) and to regulate *MAPK1*, which was found in the original Lu et al. study and is also involved in synaptic transmission. *APBA1* is a neuronal adaptor protein that interacts with, stabilizes and inhibits the proteolysis of the Alzheimer's disease amyloid precursor protein (*APP*) (Jacobs et al. 2006). A large number of genes involved in neurological disorders are present in the 207 genes of our expanded gene list which were not identified in the original 463 aging genes identified by Lu et al. These include interesting genes implicated in age specific neurological diseases, such as *CDK5*, *CDK5R1*, *BCL2L1* and *PLAT* (Luetjens et al. 2001; Melchor et al. 2003; Rademakers et al. 2005; Jacobs et al. 2006). The full gene list is found in Supplementary Table 1.

The Aging Human Kidney

The third data set considered gene expression in human kidney tissue across two regions (cortex and medulla) in multiple patients of different ages (Rodwell et al. 2004). The normalized data can be downloaded from the Stanford Microarray Database. In addition to the gene expression data, there were numerous clinical parameters available on a majority of the patients. The clinical parameters included

indicators of renal pathology, such as the degree of glomerular sclerosis or arterial intimal hyalinosis (AIH). There was also information about creatinine levels, medical history and systolic and diastolic blood pressures. All of this information was used as independent variables to predict gene expression patterns. We restricted our analysis to patient samples for which there was clinical data available for all the parameters. This limited the analysis to 63 samples. First, we analyzed the Pearson correlation derived distance matrix (equation 5) for genes that had to be scored as "Present" in at least one of the samples. Three variables were found to significantly contribute to the variation in the gene expression similarity/dissimilarity: tissue type (cortex or medulla *P*-value < 0.001; proportion = 0.18), arterial intimal hyalinosis (AIH) (*P*-value < 0.001, proportion = 0.05) and past medical history (*P*-value = 0.033, proportion = 0.03). Past medical history included a history of hypertension, diabetes mellitus, chronic renal insufficiency, hepatitis B virus, hepatitis C virus or combinations of those diseases. Next, we analyzed the distance matrix computed from only the age associated genes that the authors had identified by using a linear regression model (985 genes). Interestingly, age was not a predictor of the gene expression pattern in the age associated gene expression correlation matrix. The most significant predictors were AIH (*P*-value < 0.001; proportion = 0.15), tissue type (*P*-value < 0.002; proportion = 0.08), and race (*P*-value < 0.001 proportion = 0.07) (see Supplementary Figure 3). Age was insignificant with a *P*-value of 0.63 and a contribution of less than 0.01. However, the chronicity index, which was an index developed by the authors that scores the morphological and physiological state of the kidney and was designed

to give a physiological age to the kidney, was almost significant with a *P*-value of 0.055 and proportion of variation explained of 0.02. Thus, collinearity among chronicity index and other independent variables may have prevented it from entering into the final model as a significant predictor. When we tested the independent variables individually, the chronicity index was significant (*P*-value < 0.001; proportion = 0.15).

Beyond testing whole sets of genes, the method can test specific sub-sets of genes for which it may be hypothesized that gene expression is specifically altered. For example, one may be interested in whether genes involved in the Pharm-GKB derived ACE-inhibitor pathway show altered gene expression patterns consistent with a specific form of renal pathology. Testing all the ACE-inhibitor pathway genes using the proposed procedure, we discovered that not only are there large tissue differences between the cortex and medulla of the kidney in the ACE-inhibitor pathway (*P*-value < 0.001 proportion = 0.12), but there is a significant association above tissue differences in regards to the patient's level of tubular atrophy / interstitial fibrosis (*P*-value < 0.007 proportion = 0.08, cumulative proportion = 0.20).

Evaluating Different Distance Measures

We considered the effect of the use of different distance measures on the tests for association. Although not an exhaustive study, we present this to showcase the importance of choosing a distance matrix. The choice of a distance matrix is important in a number of related contexts, such as the choice of a distance matrix for graphically

representing data in heat map or tree diagrams or in cluster analysis settings (Hughes et al. 2004; Kibbey and Calvet 2005; Trooskens et al. 2005). We re-evaluated the associations involving the above datasets using the Pearson correlation coefficient, the Spearman correlation, the Kendall Tau correlation, Lin's concordance correlation, the Euclidean distance and the Chebychev distance to derive the distance matrix (see Supplementary Table 2). This analysis considered the distance matrices constructed from the same genes used in the analyses above. Lin's concordance correlation, the Euclidean distance and the Chebychev distance emphasize the actual proximity of the numerical values of the genes used to compute the distance matrix, and hence stand in contrast to the correlation coefficient which merely considers the linear relationship of the values across the genes used (Lin 1989, Lin 2000). The choice of a distance measure influences the proportion of variation in the distance matrix explained but not necessarily the significance of the relationship between the predictor variables and the distance matrix entries. A more thorough evaluation of this issue is forthcoming by the authors.

Signal Strength and Distance Matrix

Since it is unlikely that all of the genes considered in a study will be related to a particular predictor variable, the formation of a distance matrix with all the genes may not show a signal, or as strong a signal, with the predictor variable as a distance matrix constructed with only those genes that are relevant to the predictor variable. Unfortunately, it will be difficult to know *a priori* which genes should go into the

construction of the distance matrix. Although our procedure can be used to test each gene individually, or subsets of genes, as noted, we have also considered the more 'omnibus' hypothesis testing situation in which one is interested in knowing if there is any relationship between a predictor variable and gene expression patterns as a whole or across all genes assayed in a study. We were therefore interested in determining how strong the relationship between gene expression similarity and predictor variables considered in our examples is as a function of the number of genes considered in the construction of the distance matrix. This would provide us with insight into the amount of 'noise' that could be tolerated and still allow the 'signal' relating the gene expression values and the predictor variable to appear. We therefore considered the inclusion of random, simulated gene expression values in the construction of the distance matrix, knowing that these random simulated gene expression values would saturate the signal if enough were added. Supplementary Figure 4 shows the relationship between the *F*-statistic, the proportion of variation in similarity/dissimilarity explained, and the permutation *P*-value as a function of the number of extraneous gene expression values that are added in the construction of the distance matrix for all data sets tested. Large amounts of noise reduce the overall proportion of variation in similarity/dissimilarity explained as well as the *F*-statistic, as one would expect, however the permutation test derived *P*-values remain significant. Thus, it takes the addition of approximately 98% noise to saturate the signal to the point of statistical insignificance.

Discussion

Our proposed method of analysis can easily complement many traditional and alternative methods of analysis for high-dimensional data. In fact, since the proposed procedure can be used to analyze each gene in a univariate manner, it extends traditional univariate procedures. In addition, unlike other approaches, the proposed approach does not require a reduction of the data via principal components (Yeung and Ruzzo 2001), cluster (Eisen et al. 1998), factor (Kustra et al. 2006), or multidimensional scaling analysis (Taguchi and Oono 2005). The proposed analysis procedure also differs from related procedures, such as GSEA and globalTest (Subramanian et al. 2005; Goeman et al. 2005), in that it can be used to emphasize the multivariate nature of the expression values of many genes in the same pathway and treats the system being interrogated as a whole and does not consider each individual gene in a univariate analysis which then considers the result of the univariate analyses in aggregate. The exploitation of this fact can have disadvantages, obviously, since one may be interested in knowing which particular sets of genes are the most perturbed in a particular setting. However, it is arguable that physiological perturbations and variations are likely to "re-set" the coordinated expression patterns of many genes in order to reach biochemical or physiological homeostasis or equilibrium. Thus, the assessment of the similarity of global gene expression profiles of multiple samples with different features or exposures is appropriate.

Depending on the number of data points that are selected for analysis, it is possible to over fit the regression and identify significant predictor variables whose effect could be assigned to a large number of data points, when in fact only a smaller subset of the data points is truly associated with the predictor variable. However, since the multivariate regression technique can be reduced to a univariate analysis which focuses on single data points, it is possible to identify specific subsets of the data within a larger group for which the predictor variable is having the largest significant effect. The method we have proposed is, in fact, flexible enough to be used in settings for which insight into the effects of single genes or subsets of genes is the goal. Alternatively, one could test subsets or groups of genes based on some (*a priori*) grouping factor, such as participation in a biochemical pathway or genetic network. One could also combine the proposed approach with standard non-distance-based univariate and/or cluster analysis methods to assess the significance of groups of genes identified with these methods with respect to a predictor variable. Finally, beyond testing the relevance of specific clinical or phenotypic predictor variables, the method can be used as a quality control measure to identify potential sources of non-biological error, such as technician, chip lot or dissection error. These sources of non-biological variation can be included in the multiple regression as additional independent variables and thus these factors can be controlled for in the analysis.

There are some limitations to the proposed method that go beyond the choice of a distance metric or the manner in which individual genes or groups of genes are tested. For example, it may be the case that the actual correlation patterns among

genes differ across particular groups or levels of a quantitative predictor variable. In fact, differences in correlations among the expression levels of genes across, e.g., individuals treated with different drugs, different strains of mice, older and younger individuals, etc., may reflect actual perturbations in genetic networks, possibly more so than simple differences in the achieved levels of gene expression themselves. The assumption that a single distance metric, and hence distance matrix, characterizes similarities among individuals with respect to gene expression values may be obscuring important differences among the individuals, although the degree to which this is the case is an open question. There are methods for assessing 'heteroscedasticity,' – differences in the covariances among groups of genes – across groups of individuals (Johnson and Wichern 1992; Anderson 1984; Krzanowski 1993; see also Schork et al. unpublished web-based program interface: http://polymorphism.ucsd.edu/mama/), but their application to high-dimensional data has not been pursued to the degree that analyses considering differences in average levels of expression have.

Program Availability

The source code for this statistical method is freely available at the Biopython script central page (http://biopython.org/wiki/Scriptcentral) and is being incorporated into the Biopython library. Also, the source code and a user friendly web application are being incorporated and maintained on the Schork Laboratory website (http://polymorphism.ucsd.edu/programs.html).

Acknowledgements

The authors would like to thank Ondrej Libiger for assistance with coding the program, Charles Abney for assistance in web development and Dr. Marti Anderson for advice and encouragement. NJS is supported in part by: The NHLBI Family Blood Pressure Program (FBPP; U01 HL064777-06); The NIA Longevity Consortium (U19 AG023122-01); the NIMH Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571-02); NIH R01s: HL074730-02 and HL070137-01; the UCSD Moores Cancer Center, and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator). This chapter appears as a reformatted version of the following published material:

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A*. 2006. 103:19430-5.

Figure 7. Power of MDMR

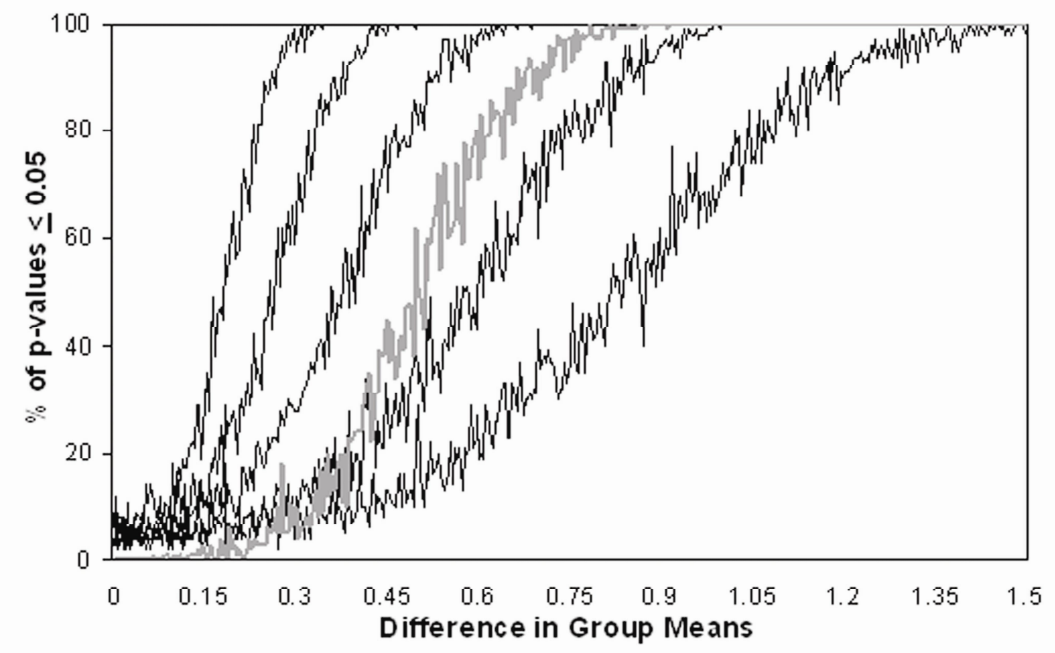


Figure 7. Power of MDMR. Power of the proposed distance matrix-based regression procedure as a function of both increasing differences in control vs. experimental settings as well as overall signal to noise ratio based simulated data sets (See text for details).

Figure 8. Optimal set with MDMR. Use of univariate regressions to identify optimal set of genes for multivariate matrix regression and hierarchical clustering. (A) Hierarchical cluster of top 100 genes identified from univariate regression of single genes. The dendrogram above the heat map shows the samples ordered from youngest to oldest (from left to right) with the leftmost branches of the tree connecting data on individuals with an average age of 43, while the rightmost branches connect data on individuals with an average age of 80 (t-test p-value < 0.000001). Hierarchical clustering was performed using CLUSTER with the average linkage metric and displayed with JAVATREEVIEW. (B) Plot of permuted p-values of age effects (purple) and sex effects (red) as well as the p-values obtained from standard ANOVA *F*-statistic for age effects (pink) and sex (green) effects. (C) Genes were ranked by the age *F*-statistic and then grouped together to identify the set of highly significant age-related genes (here found to be 100 genes) whose expression levels would produce a sample-based distance matrix such that the variation in its elements could be explained maximally by age effects. (D) A heat map matrix of the optimal set of 100 highly significant age-related genes for which age explains approximately 52% of the variation in dissimilarity across the individuals based on these genes' expression values.

Figure 8. Optimal set with MDMR

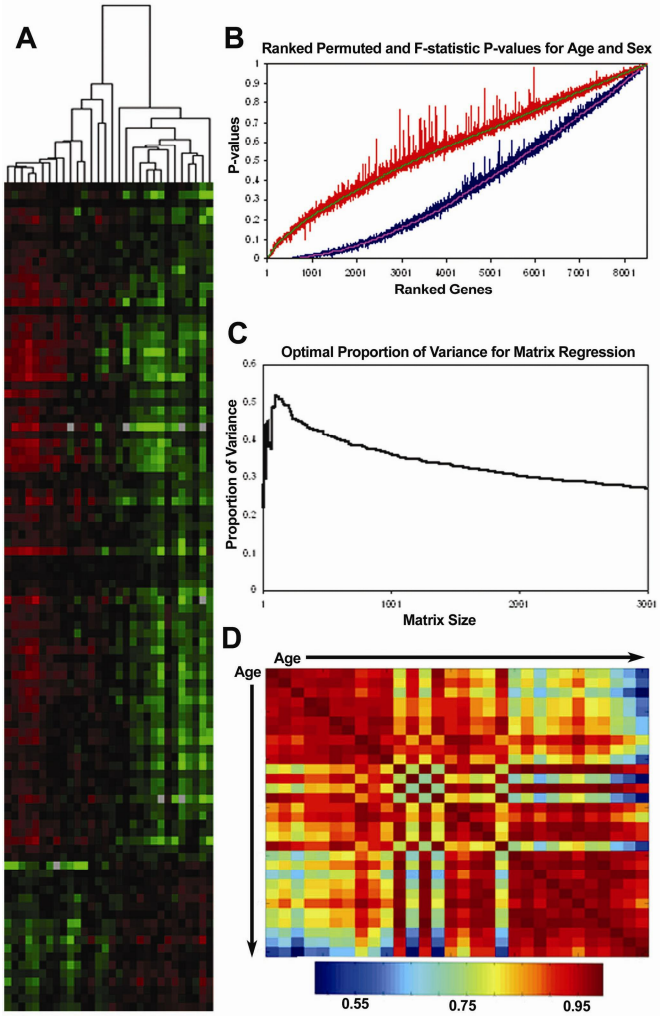


Table 4. Level Accuracy of the Proposed Permutation Test

P-values were calculated using the proposed method based on 1000 permutations of the data. The simulations were repeated 1000 times and the percentage of *P*-values below a certain threshold is reported for each of the following metrics: Pearson Correlation (Pear.), Spearman Rank (Spear.), Kendall Tau (Kend.), Lin's Concordance Correlation (Conc.), Euclidean Distance (Eucl.) and Chebychev distance (Cheb).

Distance Metric	% of tests P-value ≤ 0.01	% of tests P-value ≤ 0.05	% of tests P-value ≤ 0.25	% of tests P-value ≤ 0.50
Pear.	1.3%	4.8%	26.5%	51.4%
Spear.	1.5%	4.6%	27.4%	52.9%
Kend.	1.3%	4.9%	23.9%	47.9%
Conc.	1.3%	4.8%	26.7%	52%
Eucl.	1.2%	6.1%	24.7%	49.1%
Cheb.	1.3%	5.9%	25.4%	48.6%

References

Allison DB, Cui X, Page GP, Sabripour M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7:55-65.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* 96:6745-50.

Alter O, Brown PO, Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101-6.

Alter O, Brown PO, Botstein D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A* 100:3351-6.

Anderson MJ. (2001). A new method for non-parametric multivariate analysis of variance. *Australian Ecology* 26:32-46.

Anderson MJ. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62:245-53.

Anderson TW. (1984) *An Introduction to Multivariate Analysis*. John Wiley, New York.

Bassett DE, Eisen MB, Boguski MS. (1999). Gene expression informatics--it's all in your mine. *Nat Genet* 21:51-5.

Bryan, J. (2004). Problems in gene clustering based on gene expression data. Journal of Multivariate Analysis 90:44-66.

D'Haeseleer P. (2005). How does gene expression clustering work? *Nat Biotechnol* 23:1499-501.

Edgington ES. (1995). *Randomization Tests*. Marcel Dekker, New York.

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-8.

Goeman JJ, Oosting J, Cleton-Jansen AM, Anninga JK, van Houwelingen HC. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* 21:1950-7.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-7.

Gomez Del Pulgar T, De Ceballos ML, Guzman M, Velasco G. (2002). Cannabinoids protect astrocytes from ceramide-induced apoptosis through the phosphatidylinositol 3-kinase/protein kinase B pathway. *J Biol Chem* 277:36527-33.

Good PI. (2000). *Permutation Tests*. Springer, New York.

Gower JC, Legendre P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* 3:1-48.

Gower JC, Krzanowski WJ. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Applied Statistics* 48:505-19.

Gygi S P, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17:994-9.

Hughes T, Hyun Y, Liberles DA. (2004). Visualising very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* 5:48.

Jacobs EH, Williams RJ, Francis PT. (2006). Cyclin-dependent kinase 5, Munc18a and Munc18-interacting protein 1/X11alpha protein up-regulation in Alzheimer's disease. *Neuroscience* 138:511-22.

Johnson RA, Wichern DW. (1992). *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey.

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.

Krzanowski WJ. (1993). Permutational tests for correlation matrices. *Statistics and Computing* 3:37-44.

Krzanowski WJ. (2002). Multifactorial analysis of distance in studies of ecological community structure. *Journal of Agricultural, Biological, and Environmental Statistics* 7:222-32.

Kustra R, Shioda R, Zhu M. (2006). A factor analysis model for functional genomics. *BMC Bioinformatics* 7:216.

Legendre P, Anderson MJ. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69:1-24.

Lin LI. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255-68.

Lin LI. (2000). Corrections. *Biometrics* 56:324–5.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675-80.

Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, Yankner BA. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature* 429:883-91.

Luetjens CM, Lankiewicz S, Bui NT, Krohn AJ, Poppe M, Prehn JH. (2001). Up-regulation of Bcl-xL in response to subtoxic beta-amyloid: role in neuronal resistance against apoptotic and oxidative injury. *Neuroscience* 102:139-50.

Manly B. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman and Hall, London.

McArdle BH, Anderson MJ. (2001). Fitting multivariate models to semi-metric distances: a comment on distance-based redundancy analysis. *Ecology* 82:290-7.

Melchor JP, Pawlak R, Strickland S. (2003). The tissue plasminogen activator-plasminogen proteolytic cascade accelerates amyloid-beta (Abeta) degradation and inhibits Abeta-induced neurodegeneration. *J Neurosci* 23:8867-71.

Neter J, Wasserman W, Kutner MH. (1985) *Applied Linear Statistical Models*. Richard D. Irwin, Inc., Illinois.

Rademakers R, Sleegers K, Theuns J, Van den Broeck M, Bel Kacem S, Nilsson LG, Adolfsson R, van Duijn CM, Van Broeckhoven C, Cruts M. (2005).

Association of cyclin-dependent kinase 5 and neuronal activators p35 and p39 complex in early-onset Alzheimer's disease. *Neurobiol Aging* 26:1145-51.

Rodwell GE, Sonu R, Zahn JM, Lund J, Wilhelmy J, Wang L, Xiao W, Mindrinos M, Crane E, Segal E, Myers BD, Brooks JD, Davis RW, Higgins J, Owen AB, Kim SK. (2004). A transcriptional profile of aging in the human kidney. *PLoS Biol* 2:e427.

Slonim DK. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 32:502-8.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9:3273-97.

Storey JD, Tibshirani R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100:9440-5.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. Mesirov JP. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-50.

Taguchi YH, Oono Y. (2005). Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics* 21:730-40.

Trooskens G, De Beule D, Decouttere F, Van Criekinge W. (2005). Phylogenetic trees: visualizing, customizing and detecting incongruence. *Bioinformatics* 21:3801-2.

Webb AR. (2002). *Statistical Pattern Recognition*. John Wiley and Sons Ltd., Chichester.

Yeung KY, Ruzzo WL. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* 17:763-74.

Zapala MA, Hovatta I, Ellison JA, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Lockhart DJ, Reilly JF, Young WG, Bloom FE, Lockhart DJ, Barlow C. (2005). Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A* 102:10357-62.

CHAPTER 6

Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis

Abstract

Multivariate distance matrix regression (MDMR) analysis is a statistical technique that allows researchers to relate *P* variables to an additional *M* factors collected on *N* individuals. The technique can be applied to a number of research settings involving high dimensional data types such as DNA sequence data, gene expression microarray data and imaging data. MDMR analysis involves computing the distance between all pairs of individuals with respect to the high dimensional data of interest and constructing an *N* x *N* matrix whose elements reflect these distances. Permutation tests can be used to test linear hypothesis that consider whether or not the *M* additional factors collected on the individuals can explain variation in the observed distances between and among the individuals reflected in the matrix. MDMR analysis is an excellent complement to cluster analysis and other traditional multivariate analysis techniques. Despite its appeal and utility, properties of the statistics used in MDMR analysis have not been explored in detail. In this paper we consider the level accuracy and power of MDMR analysis assuming different distance measures and analysis settings. We also describe the utility of MDMR analysis in assessing hypotheses about the appropriate number of clusters arising from a cluster analysis.

Introduction

Contemporary biological research has become increasingly data and information intensive. Technologies such as high-throughput DNA sequencing and genotyping platforms, gene expression microarrays, imaging technologies, and continuous clinical monitoring devices provide researchers with an unprecedented amount of data for individual investigations. As a result, appropriate multivariate data analysis methods are necessary in order to test specific hypotheses or extract meaningful patterns from the data generated by these technologies. Unfortunately, many traditional data analysis procedures are not immediately applicable to high-dimensional data sets and research of the type that exploits these technologies. The reasons for this are somewhat obvious in that most traditional statistical methods were designed to test very specific hypotheses in settings for which the sample size, N, is much greater than the number of variables, P, collected on the individuals used to test the relevant hypotheses (i.e., $N>>P$).

DNA sequencing, microarray, imaging, and related studies typically generate huge amounts of data that, due to their expense and sophistication, are often collected on a relatively small number of individuals. Thus, it is typically the case that $P>>N$ in these studies. In these settings, standard univariate focused-hypothesis testing strategies are inappropriate, and their naïve application could potentially generate an enormous number of false positive findings. As an alternative to classical univariate procedures – as well as multivariate procedures designed for use with a small number of variables (such as MANOVA and multivariate regression analysis) – many researchers have resorted to analysis strategies that consider some form of data

reduction, such as cluster analysis and factor analysis (Alter et al. 2000; Quackenbush, 2001).

Although data reduction strategies have yielded important insights and have continually been refined, they do suffer from at least four problems. First, there are a myriad of different strategies for cluster analysis (such as hierarchical clustering (Eisen et al. 1998), k-means clustering (Tavazoie et al. 1999), self-organizing maps (Tamayo et al. 1999), etc.) and related strategies, making it difficult to know which approach might be the most appropriate for a given situation. Second, it is often difficult to determine, with some confidence, just how many clusters, eigenvalues, principal components, latent factors, etc. underlie or best represent any given data set. Third, the generalizability of the clusters or principal components identified from a data set, as well as their ultimate biological meaning, is often in doubt. Lastly, many data reduction procedures focus on the initial "reduction" of the dimensions of the data into a few clusters, principal components, or latent factors, and do not necessarily provide a means for drawing probabilistic inferences about the relationships of the high-dimensional data to ancillary variables of interest which, in fact, may have motivated the study in the first place. Thus, for example, one may be interested in relating tumor gene expressions patterns gathered on a set of patients to their survival or other clinical outcomes. Although one could identify clusters of patients based on their tumor gene expression profiles and test to see if the patients in those clusters exhibited different survival rates, such approaches tend to be *ad hoc* and raise the issues described in the first three problems.

We have been developing an alternative and complementary data analysis approach to data reduction procedures that does not rely on – but could still exploit aspects of – data reduction strategies. This approach, termed Multivariate Distance Matrix Regression (MDMR) analysis, is rooted in traditional linear models and was first briefly proposed in the literature by McArdle and Anderson (2001). MDMR provides a method for testing the association between a set of ancillary "independent" variables, such as a clinical outcome in a tumor gene expression study, and high dimensional data of the type produced by modern high-throughput biological assays. MDMR considers the data arising from a high-dimensional assay as providing a multivariate profile of each individual in the study. The similarity and differences in these profiles are then used to construct a distance or dissimilarity matrix whose elements are then tested for association with ancillary variables of interest. Thus, MDMR is not unlike many data reduction strategies in that it requires a distance matrix. However, unlike data reduction strategies, MDMR tests the association between the elements of the distance or dissimilarity matrix directly with the ancillary variables and therefore does not require the problematic data reduction step. MDMR can be used with all the variables resulting from a high-throughput biological assay or some subset, making it a flexible and attractive tool for identifying meaningful patterns in high-dimensional data sets.

We have described applications of MDMR to actual biological data analysis settings involving genotype data (Wessel and Schork, 2006) and gene expression data (Zapala and Schork, 2006). However, to date there has not been a study investigating

the properties of the MDMR procedure, including relevant test statistic distributions, the power of MDMR, and the robustness of the procedure. In the following, we examine the properties of the test statistics used in MDMR analysis in a wide variety of settings. We find that the MDMR test statistics and the procedure as a whole have some very desirable properties, such as an intuitive number of degrees of freedom for use in assessing the distribution of appropriate test statistics, an excellent test level accuracy, good power, and a flexibility that will make it an excellent adjunct or alternative to data reduction-based multivariate analysis strategies.

Methods

We describe the MDMR analysis procedure by considering different aspects of its formulation and properties. We note that although graphical displays of distance matrices are not an essential ingredient of MDMR analysis, we include a discussion of graphical representations because they are used routinely in contexts for which MDMR analysis is appropriate.

Computing a Distance Matrix

The formation of an appropriate distance (or dissimilarity) matrix is an essential ingredient in MDMR analysis. However, there are a large number of potential distance measures one could use to construct this matrix (Webb 2002) and unfortunately, there is very little published material that can be used to guide a researcher as to which distance measure is the most appropriate for a given situation. For example, although the Euclidean distance measure is used routinely in traditional

cluster analysis settings, functions of the correlation coefficient are the most widely used distance measures in high-dimensional gene expression analyses (D'Haeseleer 2005). We note that distance measures with either metric or non-metric properties can be used in the MDMR analyses (Gower and Krzanowski, 1999). Assuming that one has identified an appropriate distance measure, an *N x N* distance matrix is constructed. Let this distance matrix and its elements be denoted by $D = d_{ij}$ *(i,j=1,...,N)*.

MDMR Test Statistic Derivation

Once one has computed a distance matrix, *D*, the relationship between *M* additional factors (i.e. predictor or regressor variables) collected on the individuals (e.g., diagnosis, age, gender, blood pressure level, etc.) and variation in the distances between and among the *N* individuals represented in *D* can be explored. Let X be an *N x M* matrix harboring information on the *M* factors which will be modeled as the predictor or regressor variables whose relationships to the values in the distance matrix are of interest. Compute the standard projection matrix, $H = X(X'X)^{-1}X'$, typically used to estimate coefficients relating predictor variables to outcome variables in multiple regression contexts. Next, compute the matrix $A = (a_{ij}) = (-[1/2]d_{ij}^2)$ and center this matrix using the transformation discussed by Gower (Gower 1966) and denote this matrix *G*:

$$G = \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}'\right)A\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}'\right) \qquad (1)$$

where **1** is an *N*-dimensional vector of 1's. An *F*-statistic can be constructed to test the hypothesis that the *M* regressor variables have no relationship to variation in the distance or dissimilarity of the *N* subjects reflected in the *N* x *N* distance/dissimilarity matrix as (McArdle and Anderson, 2001):

$$F = \frac{tr(HGH)}{tr[(I-H)G(I-H)]} \qquad (2)$$

If the Euclidean distance is used to construct the distance matrix on a single quantitative variable (i.e. $P = 1$, as in a univariate analysis of that variable) and appropriate numerator and denominator degrees-of-freedom are accommodated in the test statistics, the *F*-statistic above is equivalent to the standard ANOVA *F*-statistic (McArdle and Anderson, 2001). The appropriate number and degrees of freedom to use in assessing significance of the test statistic in situations involving multiple variables ($P>1$) and non-Euclidean distances measures is one of the main items to be explored in the studies described in the Results section (section 3) below.

Collinearity

A fundamental problem with all multiple regression based analysis techniques is collinearity or strong dependencies (i.e., correlations) among the regressor variables. Collinearity can create problems in the computation of the projection matrix $H = X(X'X)^{-1}X'$ as well as result in unstable parameter estimates. Although there are

procedures that can be used to overcome this problem, such as ridge regression and principal components regression (Mason and Perreault, 1991), we have taken advantage of orthogonal-triangular decomposition (Gunst, 1983) to form the projection matrix and have found that this works well within the context of MDMR analysis.

Permutation Tests

The distributional properties of the *F*-statistic would be complicated to derive analytically for different non-Euclidean-based distance measures, especially when these distance measures are computed across more than one variable. Simulation-based tests, such as permutation tests, can then be used to assess statistical significance of the pseudo *F*-statistic as alternatives to the use of tests based on the asymptotic distribution of the *F*-statistic (Edgington, 1995; Good, 2000; Manly, 1997; Joeckel, 1986). We have, however, pursued an investigation of the utility of the *F*-distribution in assessing the significance of the proposed pseudo-*F* test, as discussed in depth below. We also note that the *M* regressor variables assessed in an MDMR analysis can be tested individually or in a step-wise manner (McArdle and Anderson, 2001; Zapala and Schork, 2006).

Graphical Display of Similarity Matrices

Distance matrices of the type to be used in MDMR analysis can be represented graphically in a number of ways and these graphical techniques can facilitate interpretation of the results of MDMR analysis. Two of the most widely used

graphical representations include ‘heatmaps’ and coded ‘trees’ or dendrograms (Hughes et al. 2004; Kibbey et al. 2005; Trooskens et al. 2005). Heatmaps simply color code the elements of a similarity matrix that is derived from a distance 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. If the matrix is ordered such that individuals with similar values of one of the *M* potential regressor variables in an MDMR analysis are next to each other, then neighboring cells along the diagonal of the matrix (representing individuals with similar regressor values) will present patches of red, indicating a relationship between a regressor variable and similarity. Trees are constructed such that individuals with greater similarity (i.e., less distance) are placed next to each other (i.e., they are represented as adjacent branches of the tree) and less similar individuals are represented as branches some distance away from each other. By color coding the individual branches based on the values of a regressor variable possessed by the individuals they represent, one can see if there are patches of a certain color on neighboring branches, which would indicate that the regressor variable clusters along with similarity. Similarity matrices can be easily derived from distance matrices using appropriate transformations, such as dividing each entry in the distance matrix by the empirical or theoretical maximum distance and subtracting this value from 1.0.

Cluster Analyses Involving Distance Matrices

Many forms of cluster analyses involve the use of distance matrices, such as hierarchical clustering techniques (Krzanowski 1990). As noted in the Introduction

(section I), one particularly thorny issue in cluster analysis is the determination of the optimal or most representative number of clusters in a data set. The MDMR analysis technique has utility either as an alternative to cluster analysis or as a method for determining the optimal number of clusters. Although our motivation for introducing and assessing the properties of the MDMR method is rooted in the former, we have also considered studies that assess the utility of the MDMR in the latter setting.

Results

Test Level Accuracy

The test level accuracy for the permutation test derived p-values as a function of sample size was assessed with simulated data. 100 samples ($N = 100$) were generated each with 10 random variables ($P = 10$) following a standard normal distribution with a mean of 0 and a variance of 1. 50 samples were assigned to a control group (0) and 50 samples were assigned to an experimental group (1). 1000 simulations were generated in this setting, which thus involved a single regressor variable ($M=1$) representing group membership (i.e., coded as 0 or 1) that was not associated with the 10 variables used to construct the distance matrix. Table 1 describes the results and suggests that as the sample size decreases, the permutation-test level accuracy declines.

The level accuracy is slightly improved when continuous variables are considered as regressor variables. We generated 100 samples that had 10 random variables following a standard normal distribution with a mean of 0 and a variance of

1, as in the previous setting. A random variable with mean of 0 and variance of 1 was generated for each sample and used as a single continuous regressor variable ($M=1$). 1000 simulations in this setting were conducted. Table 2 describes the results and suggests that permutation tests involving a single continuous regressor variable tend to have better level accuracy than those involving a single dichotomous regressor variable (compare Tables 1 and 2). We note that test level accuracy assuming different distance metrics was addressed in previously published work and suggests that different distance matrices do not have an appreciable effect on the behavior of permutation tests (Zapala and Schork, 2006). In addition, we have tested the level accuracy with bimodal distributions and skewed log normal distributions (results available as Supplementary Material) and obtain similar results to the normal distribution level accuracy.

Comparison with F-statistic and F-distribution

The pseudo F-statistic defined in equation 2.2.2 has a clear relationship to the F-distribution that is based on the number of quantitative variables that go into the construction of the distance matrix as well as the sample size. For a Euclidean based distance matrix involving a single variable, the appropriate degrees of freedom are related to both the sample size and the number of variables used to create the distance matrix, as noted. This can be generalized such that if one has N subjects for which there are P quantitative variables that will be used to create the distance distance, the numerator and denominator degrees of freedom for the pseudo F-statistic will be P

and (P x N) – 2 respectively, which reduces to the appropriate degrees of freedom for the standard ANOVA. We expanded the simulation studies of the type discussed in section 3.1 to compare p-values resulting from permutation tests to those derived from the F-distribution with P and (P x N) – 2 degrees of freedom. Figure 1 and Figure 2 show a clear relationship between the psuedo F-statistic, the permutation test-derived p-values and the F-distribution derived p-values. We also investigated the correspondence of the permutation test-derived p-values and the F-distribution derived p-values for small sample sizes. Figure 3 and Table 3 provide the results of these investigations and clearly show that permutation test and F-distribution derived p-values do not agree well with samples of size 10 as opposed to 100 (Figure 1). The size of the matrix, which is related to the number of subjects, affects the accuracy of the permutation test and related F-distribution-based test.

Table 3 suggests that for samples of size 10 or less the accuracy of the F-distribution based p-values suffer; however, it is considerably more accurate than the permutation test derived p-values, (compare Table 1). Figure 4 provides a scatter plot comparing p-values obtained from permutation tests versus p-values obtained from the F-distribution for data sets with sample sizes between 4 and 100 and a random number of variables ranging from 1 to 100 for MDMR analysis settings involving a single continuous regressor variable. Figure 4 clearly shows that smaller sample sizes ($N \leq$ 8) show marked differences between the permutation test derived p-values and the F-distribution derived p-values.

Power

We also pursued simulation studies to explore the power of the MDMR procedure in a variety of settings. Our initial power studies considered 30 subjects (N = 30) with 100 variables (P = 100), where these 100 variables were generated as standard normal variates. We then added a value, in increments of 0.001, to the means of the variables for 15 of the 30 subjects and tested the association between a single dichotomous categorical regressor variable (coded as 0 for the first 15 subjects and 1 for the second 15 subjects) and the distance matrix computed from the 100 variables for each subject via the Euclidean distance measure. Figure 5 displays the results for settings in which different proportions of the 100 variables had increments of 0.001 added to them for the second 15 subjects. As can be seen, when all the variables have their means adjusted for the second 15 subjects, MDMR can detect a mean difference of 0.24 80% of the time, whereas Bonferroni corrected Student's t-tests pursued on each of the P variables individually can detect a mean difference in one of the variables of 0.62 80% of the time. We also pursued power studies where the variables followed a bimodal distribution (and found that power is the same as a single mode normal distribution), skewed log normal distributions (can detect a mean difference of 0.17 80% of the time) as well as multivariate normal distributions (can detect a correlation difference among the variables of 0.06 80% of the time). These simulation studies (available as Supplementary Material) demonstrated that the MDMR procedure has similar power to detect differences in these settings and thus suggests that the MDMR procedure is robust and can detect subtle differences in groups over a range of conditions. We also considered the power of the MDMR procedure as a

function of sample size. Figure 6 depicts the results for increasing sample size assuming different mean differences between 100 normally distributed variables in two groups. It can be seen that samples sizes greater than 40 are able to identify mean differences of 0.2 or greater 80% of the time.

Finally, we studied the power of the MDMR procedure with continuous regressor variables. We induced relationships between the continuous regressor variables and the *P* variables assigned to each subject used to construct the matrix by assuming that the regressor variable was correlated at some level with either each of these P=100 variables or some fraction of them. Figure 7 depicts the results and shows that the MDMR procedure can identify relationships among data points when 15% of variables are correlated with the regressor variables at a strength of 0.2. Higher correlations allow a smaller percentage of the variables to be correlated with the regressor before the relationships are detectable with MDMR. For situations in which one may have multiple variables (i.e., $P > 1$) we note that MDMR is flexible enough to be used in a univariate manner to analyze each variable independently ($P = 1$) and identify a subset of variables for which the regressor has the strongest association with variation in the distance matrix as a whole. MDMR can then be used in a multivariate manner to determine if the overall effect of the regressor is increased by looking at these data points together. In this way, MDMR can reduce the possibility of over-fitting data and identify optimal subsets of variables related to a set of additional factors or regressor variables.

Determining the Optimal Number of Groups in a Cluster Analysis

The MDMR analysis provides an alternative to many standard multivariate analysis techniques, including cluster analysis techniques. Cluster analysis has been a common strategy used to identify patterns in high-dimensional data sets. However, given the vast array of cluster analysis strategies that have been proposed, it is often unclear which cluster analysis method is most appropriate for a particular setting. Furthermore, cluster analysis techniques rarely provide formal statistical tests to relate predictor or regressor variables to the clusters arising from an analysis and often provide ambiguous answers to questions concerning the optimal number of clusters present in a dataset. We have compared the common UPGMA (Unweighted Pair Group Method with Arithmetic mean) hierarchical clustering technique to the MDMR procedure in a single analysis setting to showcase the potential MDMR has to complement cluster analysis strategies. We generated data for two groups of subjects of size $N = 30$, where each subject was assigned $P = 100$ variables as standard normal variates. Then for the second group of subjects, we added a value to the means of each of the 100 variables. We then pursued cluster analysis on the resulting data sets and tested to see if the number of groups identified from the cluster analysis was consistent with the number of groups producing the highest and most significant (in terms of p-value) F-statistic from the MDMR analysis, where predictor variables were created reflecting group membership and tested for association with the distance matrix. We found that for mean differences less than or equal to 0.75, UPGMA clustering has difficulty identifying two distinct groups for a sample size of 60. MDMR was shown to accurately identify mean differences of greater than 0.2 for a sample size of 60 (see

Figure 6). Figure 8 provides an example of the phenomenon where UPGMA clustering suggested that there were 5 groups with some misclassified observations, although the MDMR analysis suggested two groups were the most likely. Thus, MDMR analysis can be used to fashion tests for the optimal number of groups in a cluster analysis. We are exploring this theme further in additional work.

Discussion

Our studies suggest that the MDMR analysis procedure has exceptional promise as an adjunct or alternative to standard multivariate analysis methods for use with modern high-throughput biological assays. The MDMR procedure is ideally suited for settings in which $P>>N$ where a researcher is interested in analyzing multivariate data collected on a group of individuals as though that data were providing multivariate "profiles" of those individuals, rather than as data on a distinct set of variables requiring independent attention. Such settings are the rule, rather than the exception, in many modern biological experiments. For example, gene expression studies are typically pursued to address questions about the 'state' of a cell or tissue type at a particular time or after a particular intervention. Although there is great interest in finding particular genes whose expression levels differ the greatest between times or interventions, there is also great interest in determining if the overall expression profiles of the genes have been altered or if particular groups of genes, defined by biochemical pathways or networks, have been changed. By constructing multivariate gene expression profiles of all or subsets of the genes whose similarities and differences can be interrogated, one can test hypotheses about the overall state of

the cell or tissue. For example, we have previously shown that genes involved in Pharm-GKB derived ACE-inhibitor pathway show altered multivariate gene expression patterns in the kidneys of patients with renal disease which is consistent with their levels of tubular atrophy / interstitial fibrosis (Zapala and Schork, 2006). This analysis formally tested a well-established hypothesis, that the renin-angiotensin-aldosterone system (RAAS) plays a role in renal fibrosis (Lewis et al. 2001). This type of hypothesis could not have been tested using traditional univariate or clustering approaches. We emphasize, however, that this type of analysis is in no way limited to this hypothesis, but rather can be extended to other sets of genes.

As another example, consider modern high-throughput DNA sequence data. Such data are often generated to address questions about the evolutionary relationships between species or the divergence of individuals within a species based on events such as migration, isolation, drift and/or phenotypic divergence (Wessel and Schork 2006; Nievergelt and Schork 2007). A fundamental step in the analysis of DNA sequence data to address such questions is the derivation and use of a measure of DNA sequence similarity (Clark 2006; Phillips 2006). Once one has quantified just how similar or different various DNA sequences are, hypotheses about the factors that may be associated with the differences can be framed. MDMR analysis would be an ideal tool for testing these hypotheses, especially since one would not likely be interested in testing hypotheses about differences at each nucleotide, but rather the DNA sequence as a whole or a profile.

Our studies also show that the properties of test statistics for pursuing MDMR analysis are quite good, in that they are well-behaved, exhibit an excellent level accuracy and have good power to detect a wide-range of multivariate phenomena. In addition, by confirming that the *F*-statistic used to test associations within the MDMR framework follows an *F*-distribution with an intuitive number of degrees-of-freedom, there is a computationally efficient alternative to permutation-based tests. This computational efficiency can be of great value if MDMR analyses are to be pursued in settings where repeated tests are to be performed, such as in testing associations between hundreds of thousands of DNA sequence variations and multivariate phenotypes with a whole genome association (WGA) study.

There are a number of issues with MDMR analysis that need further attention. For example, the choice of an appropriate distance measure may be problematic. Although our experience suggests that different distance measures provide roughly the same inferences (Zapala and Schork 2006), greater research into this issue should be pursued. In addition, the handling of missing data in both the construction of the distance matrix and in relating the regressor variables to the variation in the distance matrix is problematic. Handling missing data in the construction of the distance matrix may not be a huge problem if, for any pair of individuals in the sample *P* is large and they are only missing a few value between them. In this case, one could compute the distance measure with only the non-missing values. However, studies investigating the 'critical level' of missing data that can be tolerated in this setting are needed.

What would be of greatest interest, however, is a comparison of MDMR analysis with other analysis methods that could be applied to similar types of data sets. For example, for small *P* in settings involving group comparisons, one could compare MDMR with standard MANOVA or multivariate regression analyses (as done, for example, by Waters and Cohen 2006). More interesting comparisons might involve MDMR analyses in settings where *P* is large and cluster analysis, principal components and related data reduction analysis techniques might be appropriate. Regardless of the outcomes of these proposed studies, MDMR analysis has a place in multivariate analysis as one of the few approaches to directly relate variation in a large set of variables to a set of potential explanatory variables.

The source code for this statistical method is written in Python and is freely available at the Biopython script central page (http://biopython.org/wiki/Scriptcentral) and is being incorporated into the Biopython library. Also, the source code and a user friendly web application are available on the Schork Laboratory website (http://polymorphism.scripps.edu/~cabney/cgi-bin/mmr.cgi).

Acknowledgements

NJS and his laboratory are supported in part by the following research grants: The NHLBI Family Blood Pressure Program (FBPP; U01 HL064777-06); The NIA Longevity Consortium (U19 AG023122-01); the NIMH Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571-02); NIH R01s: HL074730-02 and HL070137-01; Scripps Genomic Medicine and the Donald W. Reynolds Foundation

(Helen Hobbs, Principal Investigator). The authors would like to thank Marti J. Anderson for advice and encouragement in the use of distance matrix-based regression analysis. This chapter appears as a reformatted version of the following submitted material to be published:

Zapala MA, Schork NJ. Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis. *Communication in Statistics: Simulation and Computation*. 2007. Submitted.

Figure 9. Permutation versus F-distribution 1

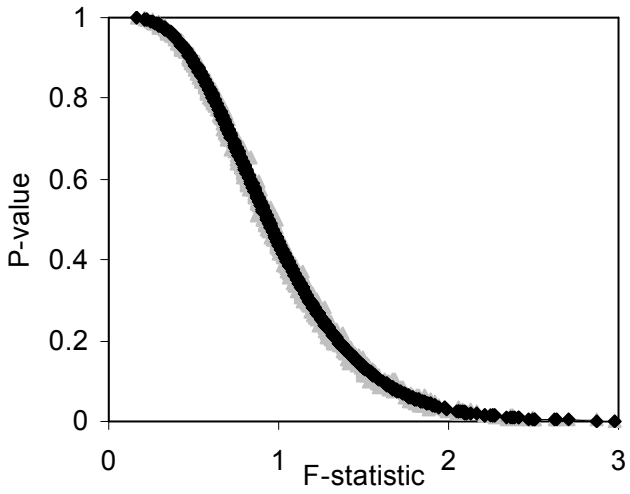


Figure 9. Permutation versus F-distribution 1. Plot of permutation test-derived p-values as a function of the *F*-statistic in gray, the corresponding p-values derived from the *F*-distribution are overlaid in black for 100 samples and 10 random variables following a normal distribution with a mean of 0 and a variance of 1 simulated 1000 times. 50 samples were coded as control (0) and 50 samples were coded as experiment (1).

Figure 10. Permutation versus F-distribution 2

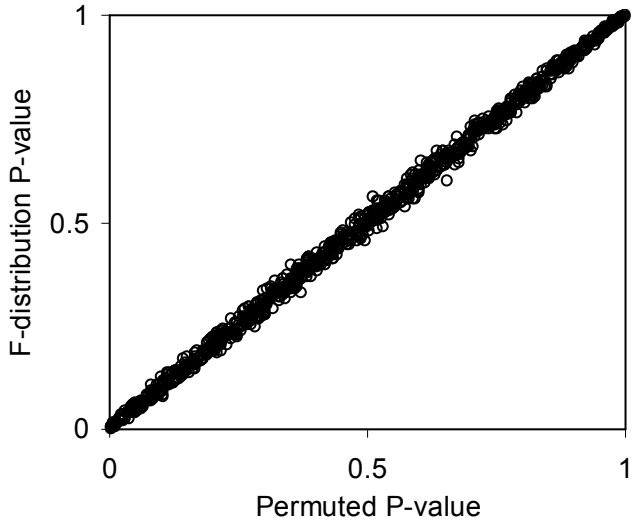


Figure 10. Permutation versus F-distribution 2. Scatter plot of p-values from Fig. 1 generated from permutation tests versus those derived from the *F*-distribution (Pearson correlation coefficient = 0.99).

Figure 11. Permutation versus F-distribution 3

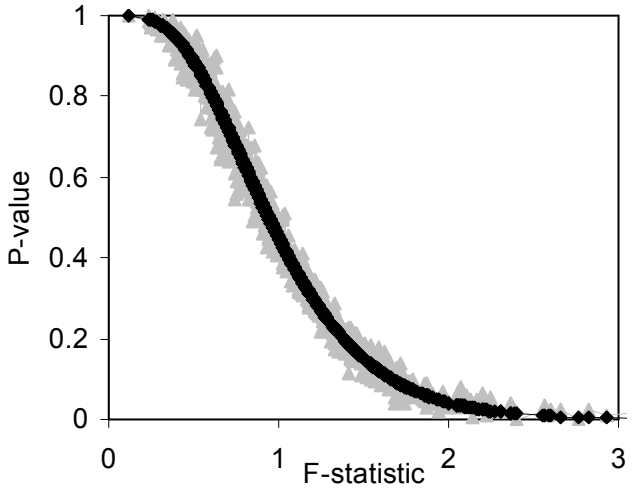


Figure 11. Permutation versus F-distribution 3. Plot of permution test-derived p-values as a function of the *F*-statistic in gray, the corresponding p-values derived from the *F*-distribution are overlaid in black for 10 samples (N=10) and 10 random variables (P=10) following a normal distribution with a mean of 0 and a variance of 1 simulated 1000 times. 5 samples were coded as control (0) and 5 samples were coded as experiment (1).

Figure 12. Permutation versus F-distribution 4

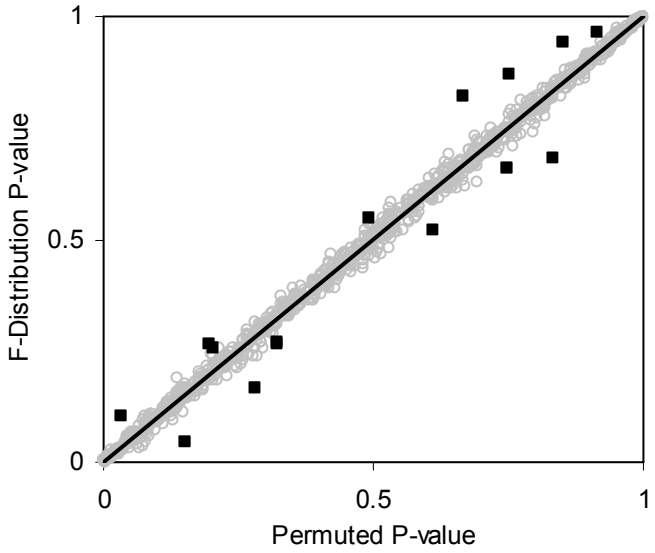


Figure 12. Permutation versus F-distribution 4. Scatter plot of p-values obtained from the *F*-distribution versus permutation tests for random samples sizes varying between 4 and 100 (i.e., $4 \leq N \leq 100$) and random variables size from 1 to 100 (i.e., $1 \leq P \leq 100$) with a single continuous regressor variable ($M = 1$) simulated 1000 times. Outlying observations represented as black squares lying away from the trend line have sample sizes less than or equal to 8.

Figure 13. Power of MDMR at % variables

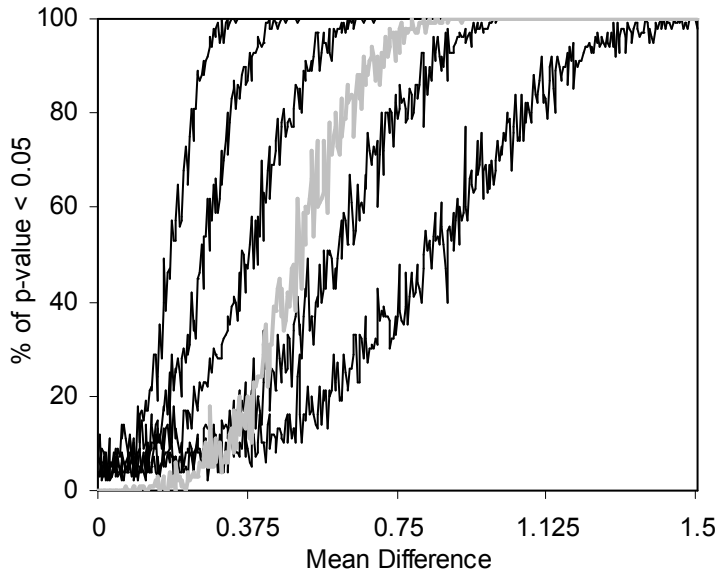


Figure 13. Power of MDMR at % variables. Power of the MDMR procedure as a function of signal-to-noise ratio obtained from 1000 simulated data sets for a wide variety of settings. Simulated data for 30 (N = 30) samples and 100 variables (P = 100) were generated with 15 samples assigned to a control group (independent variable = 0) and 15 samples assigned to an experimental group (independent variable = 1). Random data in the control group were generated as standard normal variates with a mean of 0 and variance 1. Random data in the experimental group were generated as standard normal variates with variance = 1 and means that took on values of 0 to 1.5 in increments of 0.001. The power of the permutation-based statistical test is presented. We generated different simulated data sets for which 100%, 50%, 25%, 10% or 5% of the variables used in the construction of the distance matrix had means adjusted from 0 (in the appropriate increments) in the experimental group. The gray line shows the power of a Bonferroni corrected p-value for the Student's t-tests performed on each of the 100 variables in univariate t-tests which were corrected for the hundred statistical tests pursued.

Figure 14. Power of MDMR with sample sizes

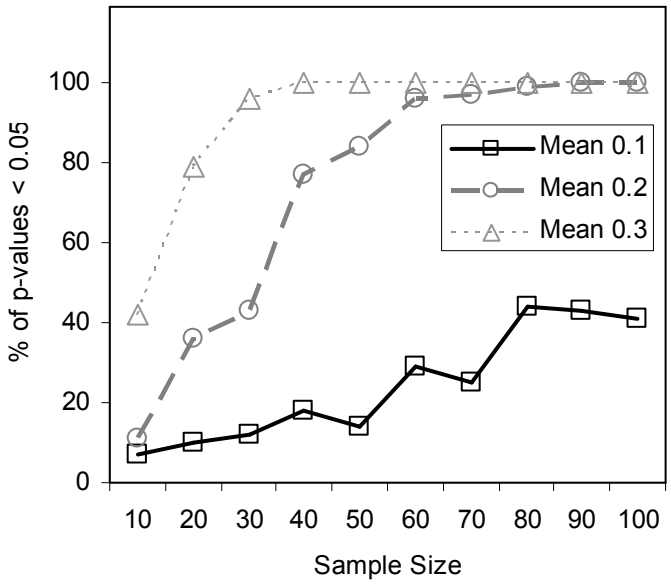


Figure 14. Power of MDMR with sample sizes. Power of the MDMR procedure as a function of increasing sample size. Half of the samples for each sample size were assigned to a control (coded as 0) and half to an experimental group (coded as 1). For each sample 100 random variables were generated following a normal distribution with a mean of 0 and a variance of 1 for the control group and an assigned mean difference of 0.1, 0.2 or 0.3 and a variance of 1 for the experimental group.

Figure 15. Power of MDMR with continuous regressor

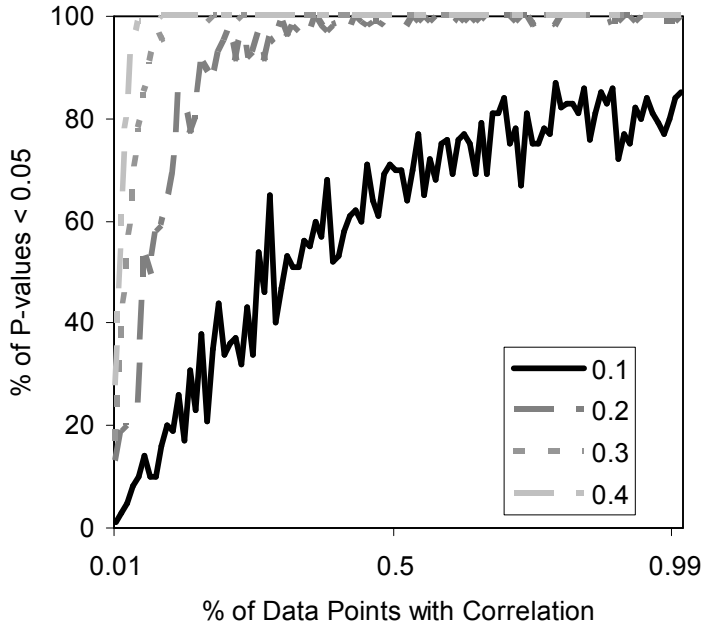


Figure 15. Power of MDMR with continuous regressor. Power of the proposed MDMR procedure as a function of the correlation of continuous regressor variables for a samples size of $N = 100$ with $P = 100$ variables. The x-axis displays the percentage of variables that have a correlation to the regressor variable. 4 different correlation strengths are shown ranging from 0.1 to 0.4. $P = 100$ random variables were generated following a normal distribution with a mean of 0 and a variance of 1.

Figure 16. Comparison of the UPGMA hierarchical cluster algorithm to MDMR

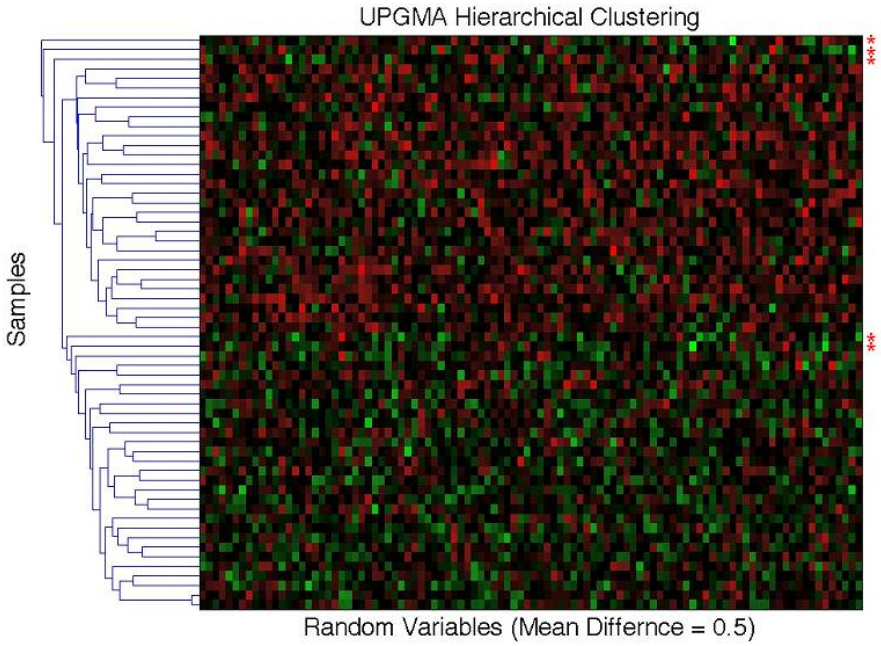


Figure 16. Comparison of the UPGMA hierarchical cluster algorithm to MDMR. Simulated data for N = 60 samples and P = 100 variables were generated with 30 samples assigned to the control group (independent variable = 0) and 30 samples assigned to the experimental group (independent variable = 1). Random data in the control group were generated as standard normal variates with a mean of 0 and variance of 1. At mean differences below 0.75, hierarchical clustering using the unweighted average distance (UPGMA) does not clearly differentiate two groups with different means. Shown above are 5 clusters for what visually appears to be two groups. The red asterisks (*) signify simulated data that has been misclassified. Two samples whose means were at 0.5 were grouped with samples whose means where 0 (bottom two asterisks). The matrix regression technique shows that the correct grouping of two separate groups gives the highest F-statistic of 5.32, while the UPGMA clustering technique of 5 distinct groups only provides an F-statistic of 5.28.

Table 5. Level accuracy of a permutation test as a function of decreasing sample size over 1000 simulations for a single dichotomous (categorical) predictor variable

	$N = 100$	$N = 50$	$N = 20$	$N = 10$	$N = 4$
1%	1.4	1.2	1.0	0.5	0.0
5%	5.8	6.4	5.1	4.9	0.3
10%	10.7	11.0	9.3	11.2	2.0
25%	25.1	24.7	29.0	25.8	10.9
50%	51.4	47.5	53.4	51.0	39.3
75%	75.8	75.1	74.8	78.4	69.5

Table 6. Level accuracy of permutations as a function of decreasing sample size over 1000 simulations for continuous variables

	$N = 100$	$N = 50$	$N = 20$	$N = 10$	$N = 4$
1%	1.4	1.5	1.2	1.6	0.0
5%	5.5	5.4	5.4	5.7	3.5
10%	10.3	11.2	11.1	12.2	7.3
25%	24.0	26.7	25.0	24.7	21.2
50%	46.6	51.3	51.3	50.7	48.1
75%	72.6	74.7	76	74.9	73.5

Table 7. Level accuracy of F-distribution p-values as a function of decreasing sample size over 1000 simulations for a single dichotomous (categorical) predictor variable

	$N = 100$	$N = 50$	$N = 20$	$N = 10$	$N = 4$
1%	1.5	0.8	1.5	1.3	2.3
5%	5.5	6.2	5.2	5.7	8.0
10%	10.5	11.3	10.4	11.0	12.8
25%	25.2	24.6	28.9	25.9	26.6
50%	51.5	46.8	53.3	52.1	52.4
75%	76.2	74.9	75.0	77.6	75.1

References

Alter O, Brown PO, Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America* 97:10101-10106.

Anderson MJ. (2001). Fitting multivariate models to community data: a comment on distance-based redundacy analysis. *Ecology* 82:290-297.

Clark AG. (2006). Genomics of the evolutionary process. *Trends in Ecological Evolution* 21:316-321.

D'Haeseleer P. (2005). How does gene expression clustering work? *Nature Biotechnology* 23:1499-1501.

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95:14863-14868.

Edgington ES. (1995). *Randomization Tests*. New York: Marcel Dekker.

Good PI. (2000). *Permutation Tests.* New York: Springer.

Gower JC. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325-338.

Gower JC, Krzanowski WJ. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Applied Statistics* 48:505-519.

Gunst RF. (1983). Regression Analysis with Multicollinear Predictor Variables: Definition, Detection and Effects. *Communications in Statistics, Theory and Methods* 12:2217–2260.

Hughes T, Hyun Y, Liberles DA. (2004). Visualising very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* 5:48.

Jockel KH. (1986). Finite sample properties and asymptotic efficiency of Monte Carl tests. *The Annals of Statistics* 14:336-347.

Kibbey C, Calvet A. (2005). Molecular Property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *Journal of Chemical Information and Modeling* 45:523-532.

Lewis EJ, Hunsicker LG, Clarke WR, Berl T, Pohl MA, Lewis JB, Ritz E, Atkins RC, Rohde R, Raz I, Collaborative Study Group. (2001). Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *New England Journal of Medicine* 345:851-860.

Manly B. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. London: Chapman and Hall.

Mason CH, Perreault WD. (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research* 28:268-280.

McArdle BH, Anderson MJ. (2001). Fitting multivariate models to semi-metric distances: a comment on distance-based redundancy analysis. *Ecology* 82:290-297.

Nievergelt CM, Libiger O, Schork NJ. (2007) Generalized Analysis of Molecular Variance. *PLoS Genetics* 3:e51.

Phillips AJ. (2006). Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics* 39:18-33.

Quackenbush J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* 2:418-427.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* 96:2907-2912.

Tavozoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. (1999). Systematic determination of genetic network architecture. *Nature Genetics* 22:281-285.

Trooskens G, De Beule D, Decouttere F, Van Criekinge W. (2005). Phylogenetic trees: visualizing, customizing and detecting incongruence. *Bioinformatics* 21:3801-3802.

Waters J, Cohen LD. (2006). A comparison of statistical approaches to analyzing community convergence between natural and constructed oyster reefs. *Journal of Experimental Marine Biology and Ecology* 330, 81-95.

Webb AR. (2002). *Statistical Pattern Recognition*. Chichester:John Wiley and Sons Ltd.

Wessel J, Schork NJ. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *American Journal of Human Genetics* 79:792-806.

Zapala MA, Schork NJ. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America* 103:19430-19435.

CHAPTER 7

High-Density QTL Mapping to Identify Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways

Abstract

Associating SNPs to gene expression levels has been useful in identifying transcriptional networks in what is called "genetical genomic" or expression Quantitative Trait Loci (eQTL) analysis. These types of studies suffer from multiple comparison problems as thousands of gene expression signals are tested against thousands of SNPs. Moreover, the biological meaning of a potential *cis* or *trans* association emerging from such studies is difficult to discern. We have taken an integrative approach using univariate, aggregate and multivariate statistics to associate gene expression values to SNPs both individually and as gene sets belonging to biochemical pathways. We analyzed public gene expression data on hematopoetic stems cells (HSC) from 22 BXD mice. 1093 unique SNPs from the Wellcome Trust SNP genotype set are available for these mice. Genes were grouped into GO, KEGG, Biocarta and Mousepath pathways based on their Entrez ID. The aggregate and multivariate analyses identified genes previously identified in univariate studies. More importantly, the aggregate and multivariate methods identified novel associations that occurred in gene sets related to hematopoetic stem cells, such as Hematopoesis, EPO Signaling, EGFR Signaling and Biosynthesis of Steroids pathways that univariate methods did not. The results presented here demonstrate the power of performing aggregate and multivariate statistical analyses in addition to univariate analysis in eQTL studies. The aggregate and multivariate analyses may allow the discovery of novel associations between genetic variation (SNPs) and expression profiles over biological pathways. Complex phenotypes are potentially

mediated by the differential expression of multiple pathways; therefore aggregate and multivariate association techniques that incorporate pathway information can be more powerful and offer more biological insight than association between SNPs and individual genes.

Introduction

The completion of the Mouse and Human HapMap projects and related DNA sequencing and polymorphism discovery initiatives have motivated researchers to consider the identification of naturally occurring DNA sequence variations that influence common complex diseases or phenotypes (International HapMap Consortium 2005; Wade and Daly 2005). The question of how to approach the problem of identifying the specific genes mediating disease susceptibility with the information provided by the HapMap and related projects is a timely but difficult one. The reason for this is that the identification, collation, and structural annotation of genes do not provide sufficient information about their functions or physiological effects to make linking genes and genetic variations to specific diseases simple. The recent convergence of dense genotype data, gene expression data and phenotype data compiled on a genomic scale could, if analyzed properly, facilitate the identification of potential disease-causing candidate loci (Chesler et al. 2005; Grubb et al. 2004).

Many groups have made interesting discoveries from the analysis of genotype and gene expression data together in what has been called “genetical genomics” or expression Quantitative Trait Loci (eQTL) studies (Jansen and Nap 2001; Schadt et al. 2003). By further incorporating well quantified phenotypes, others have discovered

genetically-mediated relationships between genotypes, gene expression and phenotypes (Schadt et al. 2005). This latter approach, in theory, could allow one to identify multiple genes that work collectively – and possibly subtly – on biochemical or physiologic processes that, when perturbed, induce disease pathogenesis.

Although genetical genomic analyses have been useful in identifying DNA sequence variations that influence transcriptional regulatory networks, they suffer from at least two potential drawbacks. The first drawback concerns the number of statistical tests performed. Since thousands of gene expression signals are typically tested against thousands of Single Nucleotide Polymorphism (SNP) markers, there is tremendous potential for false-positives. The second drawback concerns the biological meaning of a putative *cis* or *trans*-acting variation revealed by these studies. Previous genetical genomic approaches have mainly focused on treating each gene independently and ignoring the correlations between the expression patterns of genes involved in the same biochemical pathway or functional group (Grupe et al. 2001; Pletcher et al. 2004).

Public pathway and functional databases, such as KEGG (http://www.genome.jp/kegg/pathway.html) and GO (http://www.geneontology.org/), now offer the ability to group genes into biologically meaningful clusters. The ability to query these rich complimentary data sources simultaneously can provide important insights that can facilitate attempts to unravel the genetic determinants of complex polygenic diseases and phenotypes. Recently, gene expression analyses have leveraged these resources using the following general recipe:

1. Identify groups of genes that show differential expression with respect to a single phenotype of interest using parametric statistics (Hosack et al. 2003) or more robust non-parametric statistics (Goeman et al. 2005; Subramanian et al. 2005; Ye and Eskin 2007).

2. Use a dimensionality reduction technique, such as principal component analysis, to identify several principal gene expression signals that reflect mechanisms that coordinate gene expression (Schadt et al. 2005).

3. Perform traditional linkage and association analyses to identify *cis* or *trans* eQTLs (Schadt et al. 2005; Ghazalpour et al. 2006).

There are two problems with the above approach. One, the principal gene expression signals are often not biologically meaningful as they represent linear combinations of the expression levels of the genes and not the expression levels of the genes themselves. Second, reducing a complex dataset to a few representative components neglects the multivariate nature of gene expression levels associated with gene sets, such as biochemical pathways, and therefore does not consider the gene set or system being interrogated as a whole. Our approach allows a researcher to identify sequence variations that affect entire pathways or common gene functions in a multivariate statistical manner.

Approach

We have previously developed methods to improve upon univariate gene set-based methods by focusing analysis on the discovery of significant gene sets which are most biologically relevant. We formalized the notions of 'differentially expressed' and

'tightly regulated gene sets' with or without respect to a phenotype of interest. It is our belief that those gene sets exhibiting both properties (i.e., tightly regulated, differentially expressed genes that appear related to a phenotype of interest) are the most likely to represent genes of greatest biological interest (Ye and Eskin 2007). We extend the notion of differentially expressed gene sets to identify individual SNPs that are associated with gene set expression. Specifically, we perform ANOVA on each gene-SNP pair in a population (where genotyping categories are taken as the grouping or independent variable and expression levels are taken as the dependent variable) and find those SNPs most highly associated with a gene set using an aggregate statistic such as the Mann Whitney Test. In the aggregate approach, we examine all expression levels in a gene set by performing associations between each gene and SNP and reduce these associations to a single aggregate statistic. This is fundamentally different from dimensionality reduction techniques which reduce the expression profiles to a few components and then perform association using only those components.

This ANOVA based approach can be improved through the development and implementation of multivariate statistical methods that consider the expression levels of all genes in a gene set. We describe such a method that is analogous to standard ANOVA and related linear models. This analysis method considers the relationship between variables, such as phenotypes and/or SNP genotype categories, to the pair-wise similarity/distances in gene expression values represented in a matrix. Essentially, this matrix reflects the proximity of gene expression profiles between and

among the individuals with respect to the expression levels of genes in a particular gene set (Zapala and Schork 2006). The proposed multivariate method avoids the need for reducing the dimension of the similarity/distance matrix, can be used to assess relationships between the overall expression levels of the genes and any additional information collected on the samples, and can also be used to analyze individual genes or groups of genes. To showcase the advantages of this proposed multivariate approach, we have compared the results of its application to the application of the step-wise univariate and aggregate association methods discussed above.

These methods can be used in the context of eQTL mapping studies with the goal of identifying functional SNPs that are associated with the expression levels of sets of genes. To showcase and compare the methods we analyzed publicly-available gene expression data gathered on hematopoetic stem cells (HSCs) isolated from 22 recombinant inbred BXD mice. The gene expression data from the Affymetrix MG-U74Av2 array with 12,488 genes was normalized using quantile normalization and expression estimates were computed using "Robust Multi-array Average" or RMA (Irizarry et al. 2003). These mice were also genotyped on 1093 unique SNPs from the Wellcome Trust CTC mouse strain SNP genotype set after removing monomorphic and repetitive SNPs (http://www.well.ox.ac.uk/mouse/INBREDS) (Bystrykh et al. 2005). Probe sets on the array were grouped into pathways or functional specific groups from GO Biological Processes, KEGG, Biocarta and Mousepath based on their Entrez ID (Tian et al. 2005).

Results

Univariate Association Study

A simple ANOVA-based association study was first compared to the results achieved in previous analysis efforts (Bystrykh et al. 2005). The goal of the study was to identify genes whose expression pattern can be explained by a variation at a single locus (i.e., genotype categories for a given SNP). This analysis involved performing a standard ANOVA *F* test where the SNP genotypes were taken as the grouping factor. Table 1 shows the top 20 scoring (SNP, gene) pairs sorted by their observed *F* statistics. The *p*-value for each *F* statistic was calculated assuming a standard *F* distribution with degrees of freedom of $df_1 = 1$ for the SNP and $df_0 = 42$ for the strains (we have replicate expression levels for each strain). This list identifies many of the same significant genes identified in a previous linkage study as indicated in bold (Bystrykh et al. 2005). *Ctse*, the top scoring gene was the second highest scoring gene found by Bystryk, et al. The associated SNP (rs6250833) is found to reside in the 3' untranslated region of *Ctse.* However, the association occurs over a large stretch of the genome and could merely reflect the strong linkage disequilibrium (LD) among the loci genotyped in the BXD strains ranging from rs6263067 (Chromosome 1, base pair 128,995,783) to rs13476148 (Chromosome 1, base pair 144,782,263). This is expected since the SNP dataset was compacted to have on average one SNP per LD block. There are a large number of significant genes found by each method that are not confirmed by the other method. This is probably due to discrepancies in the genomic coverage of QTL markers used in the study by Bystrykh, et al. 2005 and the SNPs used in the association analysis. Finally, whereas Bystrykh, et al. reported

several *trans* associations, only one of the 20 top associations was found to exhibit a true *trans* relationship between the SNP and the gene (AI594671).

Gene Set Association Aggregate Statistic

The single gene ANOVA was extended to examine gene sets by defining an aggregate statistic based on the Mann Whitney U statistics that combines the results from single gene ANOVA analyses (see Methods). Given a ranked list L of F statistics associated with the expression values of each gene, a gene set is said to be associated with a SNP if the F statistics of the genes belonging to the gene set are ranked higher in the list, where a gene set represents a particular pathway. As noted, to account for large and correlated gene sets, the statistic was recomputed for each gene set over a permuted set of SNPs from the Wellcome dataset (1093 SNPs).

2474 (SNP, gene set) pairs were found with a nominal p-value of < 0.0009. The top scoring SNP from the univariate ANOVA, rs6250833 which includes *Ctse*, was found to be significantly associated with the GO Biological Process of B Cell Differentiation (GO: 0030183). It is interesting that *Ctse* is not a gene in the B Cell Differentiation pathway suggesting that it might interact with other genes in the pathway to influence expression. The results demonstrate that not only can the aggregate method identify the same SNPs as the univariate approach, it can also identify the association between SNPs and the expression levels of a set of genes that belong to the same biological pathway. These gene set associations are potentially much more biologically relevant than single gene associations.

To further illustrate the power of the aggregate approach, several significant SNPs not identified in the univariate ANOVA analysis where found to be highly associated with specific gene sets related to HSCs. These genes sets include the KEGG Hematopoietic cell lineage and the mousepaths Stem Cell pathways. The KEGG Hematopoietic cell lineage pathway was associated with rs6300458 which is a SNP in the Rbm26 gene, a RNA binding motif. The mousepaths Stem Cell pathway is associated with rs13459145, a SNP in the Lifr, leukemia inhibitor receptor gene.

Multivariate Association Study

Univariate association analysis results can lead to the identification of putative sequence variations that may influence changes in the expression levels of particular genes. Multivariate approaches of the type we have been developing allow one to determine if a particular sequence variation affects an entire set of genes more or less simultaneously, such as those in a biochemical pathway. One of the highest univariate associations that was also identified in the multivariate association occurs over a rather large stretch of the genome that contains the gene *Sc5d*. The largest multivariate pathway associated with SNPs in this particular region occurs with KEGG pathway 'Biosynthesis of Steroids' (*F*-statistic = 11.6, p-value $\leq 1 \times 10^{-16}$). The marker rs3696264 is in this region and is in the intronic region of Burkitt lymphoma receptor 1 (*Blr1*) at Chromosome 9, base pair 44,277,309. B-cell CLL/lymphoma 9-like (*Bcl9l*) is directly upstream of *Blr1* and appears to be within an LD block with *Blr1* (See Figure 4). An extended region of LD on Chromosome 9 from base pair 42,014,058 (rs6388711) to 45,183,190 (rs13480169) is shown to be associated with

several related gene sets from our aggregate analysis including Hemopoiesis (GO:0030097), Regulation of Cell Differentiation (GO:0045595), Regulation of Cytokine Biosynthesis (GO:0042035) and Hemopoietic of lymphoid organ development (GO:0048534). Some of the highest pathways for the other corresponding SNPs identified in the univariate ANOVA analysis were for Chemokine receptor 2, *Ccr2*, for the mousepaths Chemokines and Receptors (*F*-statistic = 21.04, p-value $\leq 1 \times 10^{-16}$) and the Biocarta pathway of selective expression of chemokine receptors during T-cell polarization (*F*-statistic = 33.76, p-value $\leq 1 \times 10^{-16}$).

SNPs identified in pathways from the multivariate method which were also identified in the univariate analysis raise important questions about their consistency. However, associations involving pathways which contained genes and SNPs not initially identified in the univariate analysis are intriguing in their own right. The first SNP identified in the multivariate regression not identified in the univariate analysis was rs6395893 and corresponded to the GO category of antigen presentation, exogenous antigen via MHC class II (GO:0042591) (*F*-statistic = 13.52, p-value $\leq 1 \times 10^{-16}$). Again, the association occurs over a rather large stretch of the genome and likely reflects strong linkage disequilibrium (LD) among the loci genotyped in the BXD strains ranging from rs13482947 (Chromosome 17, base pair 31,361,311) to rs6395893 (Chromosome 17, base pair 38,786,224). This stretch of the genome contains the H2 complex and 11 genes involved in antigen presentation via MHC class II and a larger group of genes involved in antigen presentation via MHC class I.

Additional SNPs identified in the multivariate analysis not found in the univariate include rs3701429 which corresponded to the Biocarta pathway of EGF receptor transactivation by GPCRs in cardiac hypertrophy (see Figure 4) (*F*-statistic = 12.87, p-value $\leq 1 \times 10^{-16}$). The association covers a region on Chromosome 6 from base pair 29,127,681 to base pair 30,223,416. This region includes the nuclear respiratory factor 1 (*Nrf1*), a transcription factor that controls the expression of other genes involved in heme biosynthesis. Lastly, rs13481632 was found to be associated with the EPO signaling pathway (*F*-statistic = 10.83, p-value $\leq 1 \times 10^{-16}$). The association covered a region on Chromosome 12 from base pair 105,915,746 to base pair 106,797,604. The region includes the bradykinin receptors beta 1 and beta 2 (*Bdkrb1* and *Bdkrb2*) which play integral roles in the inflammatory response.

Discussion

As genomic technologies are developed that focus on the high-throughput interrogation of the genome at many different levels, such as identifying sequence differences and quantifying mRNA levels associated with genes expressed in particular cells or tissues, the need for analysis techniques that integrate the information they produce in meaningful ways becomes paramount. Current genetical genomics techniques provide methodologies for discovering associations between an individual gene and a SNP that reflect the influence of *cis* or in *trans* acting factors but do not have the ability to reveal the workings of specific biological processes that are affected by the SNP. By utilizing aggregate statistics and multivariate analysis, we

have shown that an understanding of the effects of genetic variation on groups of genes and pathways is possible.

Both the aggregate analysis and the multivariate analysis we described were able to identify SNP associations that were both revealed and not revealed in the univariate analysis and place those associations within a biochemical process or pathway. For example, the univariate, aggregate and multivariate analyses all identified a strong association in a genomic region on chromosome 9 which included SNP marker rs6388711. The aggregate analysis suggested that variation in this region influenced several genes in sets related to HSCs including Hemopoiesis (GO:0030097), Regulation of Cell Differentiation (GO:0045595), Regulation of Cytokine Biosynthesis (GO:0042035) and Hemopoietic of lymphoid organ development (GO:0048534). The multivariate analysis identified the process as the KEGG Biosynthesis of Steroids pathway. In the region on Chromosome 9 there is a SNP (rs3696264) within *Blr1* which is also known as *CXCR5* that is known to play a critical role in the development of B-cell follicles in the mouse spleen (Ohl et al. 2003). It has also been shown that *Blr1* plays a role in regulating hematopoetic cell growth and differentiation (Battle and Yen 2002). Steroids have been shown clinically to induce differentiation of leukemic cells in children with acute promyelocytic leukemia (APL), which is a malignant disorder that results in the replacement of normal bone marrow with primitive HSCs (Hiçsönmez et al. 1993). Moreover, all-trans retinoic acid (ATRA) is used to treat APL but can result in retinoic acid syndrome (RAS) in some patients which is thought to be due to cytokine and

chemokine release from the differentiating hematopoetic cells (Gao et al. 2007). RAS is treated with the steroid dexamethasone and numerous studies have shown that steroids, such as dexamethsone, alter circulating cytokine and chemokine levels (Weinberg et al. 1992). Thus, it could be that SNPs in *Blr1*, a chemokine receptor that plays a role in the maturation of B-cells from HSCs, could alter the expression of groups of genes involved in the biosynthesis of steroids.

The true power of a gene set or pathway based approach lies in the ability to identify SNPs associated with pathways not identified from the univariate analysis. For example, the association for the GO category of antigen presentation with a large stretch of the genome on Chromosome 17 may not initially appear to be related to HSCs from which the gene expression was derived. However, in a previous traditional QTL mapping study analyzing HSCs using an intercross mapping approach in AKR and C57BL/6Ka-Thy1.1 mice, an identical locus on chromosome 17, including the H2 complex, was significantly linked to the frequency of long-term self-renewing HSCs (Morrison et al. 2002). *TNF* is also located within this region and has been shown to inhibit HSC proliferation (Bryder et al. 2001). Therefore, it could be that there are multiple genes within this region on Chromosome 17 involved in antigen presentation whose functional expression is working synergistically to affect HSCs, as previous studies had suggested. This region of LD associated with antigen presentation was only found using the multivariate analysis and not found using the aggregate analysis.

The results presented here demonstrate the power of performing univariate, aggregate and multivariate statistical analyses together to interrogate expression QTL data. The univariate approach allows for the identification of individual genes whose expression appears genetically regulated. The aggregate approach facilitates the discovery SNPs whose association is over-represented in a particular pathway. Finally, the multivariate analysis examines entire biological systems as they relate to DNA sequence variation. Thus, each analysis method offers unique and important information that could ultimately lead to new discoveries and hypotheses about genetic mechanisms involved in regulating gene expression as well as phenotypic expression. In this study we have focused on SNPs as the independent variables. However, the methods presented here can easily be extended to include phenotypic information of relevance. By associating pathways to both SNPs and phenotypes, researchers can begin to understand the genetic basis of complex phenotypes whose expression is the result of multiple genes working in concert.

Methods

Univariate Association Study - ANOVA

When interest is in the analysis of single SNPs, linear fixed effects Analysis of Variance (ANOVA) models can be applied to each locus. The results of these analyses can then be ranked based on the resulting *F* statistics. The null hypothesis is that genotypic categories associated with any particular SNP do not explain the observed variation in the expression values of the gene. We have used the "MAANOVA" module implemented in the R software suite to conduct such analyses and specify the

gene-specific model as: $r_{igr} = G + S_i + \varepsilon_{ir}$ where G is the average intensity associated with a particular gene, S is the effect associated with the SNP (i.e. genotypic categories or, in the context of certain genetic systems, major versus minor allele), and ε is the residual. We compute the following F statistic:

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1 / df_1} \quad (1)$$

where rss_0, df_0 and rss_1, df_1 are the residual sums of squares and degrees of freedom for the null and alternative models, respectively. A p-value based on the normal distribution is returned by MAANOVA (Cui and Churchill 2003).

Aggregate Association Study – Mann-Whitney Tests

To compute over-representation of a gene set from ranked statistics derived from linear ANOVA models, the Mann Whitney statistic was used. This non-parametric test uses the entire list of ranked F-statistics to compute a score, which provides an advantage over Fisher's exact test or the χ^2 test, which only focus on a portion of the list above a predefined threshold. Given a ranked list of genes, the alternative hypothesis is that gene sets most associated with a SNP will have a large number of members clustered at the top of the list.

Formally, given a microarray experiment involving the measurement of the expression levels of P genes, the genes are first ranked based on their F statistics derived from an ANOVA for each SNP. Each gene set of interest, denoted S, has the Mann Whitney score computed where the ranked F statistics associated with genes belonging to the set are compared to ranked F statistics associated with genes not in

the set. The alternate hypothesis is that those genes belonging to the gene set tend to be ranked higher on the list than other genes. The null hypothesis is that genes belonging to the gene set should be ranked randomly with respect to all other genes. The R statistics package was used to compute the Mann Whitney statistic. Because the Mann Whitney score is sensitive to correlation structures observed in the gene sets (i.e., the rankings of the *F* statistics associated with the genes are not independent given correlations between the expression levels of genes), the significance is computed by permuting each of the 1093 SNPs and recalculating the Mann Whitney score. Each of the original scores are compared to the 1093 new scores to achieve a nominal *p*-value of <0.0009.

Multivariate Association Study – Distance Matrix Regression

As an alternative to considering the analysis of the expression values of each single gene in isolation and then determining whether the expression values are associated with genotypic categories of a SNP, a multivariate distance matrix regression technique was recently developed that produces an *F* statistic encompassing the effect of groups of genes (Zapala and Schork 2006). For example, if one has collected *N* total individuals and assayed the expression levels of *P* genes on those individuals, then an *N* x *N* similarity or distance matrix can be formed that reflects the Euclidean distance in expression values, over the *P* gene expression levels, between each pair of individuals in the study.

Formally, let **Y** be an *N* x *P* matrix harboring gene expression values on *N* individuals for *P* genes. Let **X** be an *N* x *M* matrix harboring information on *M*

predictor or regressor variables whose relationship to the gene expression values is of interest. The predictor variables in this case contain the SNP alleles. Let **D** be an *N* x *N* distance matrix, whose elements, d_{ij}, reflect the distance (or dissimilarity) of subjects *i* and *j* with respect to the *P* gene expression values. Let $\mathbf{A} = (a_{ij}) = (-[1/2]d_{ij}^{2})$, then one can form Gower's centered matrix **G** from **A** by calculating:

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \tag{2}$$

where **1** is a *N*-dimensional column vector whose every element is 1 and **I** is an *N* x *N* identity matrix. An appropriate *F* statistic for assessing the relationship between the *M* predictor variables and variation in the dissimilarities among the *N* subjects with respect to the *P* variables (e.g., gene expression values) is:

$$F = \frac{tr(\mathbf{HGH})/(M-1)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-M)} \tag{3}$$

Where **H** is the indempotent "hat" matrix ($\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$) and **X** is an *N* x *M* matrix harboring information on *M* predictor or regressor variables (i.e. SNPs) whose relationship to the gene expression values is of interest, where the first column contains a column vector whose every element is 1, and reflects an intercept term, as in standard regression contexts. If *P* = 1 (i.e., a univariate analysis looking at the expression of a single gene) and the distance matrix is computed through the use of the standard Euclidean distance measure, then *F* in Equation 3 is the standard *F* statistic and possesses the typical properties associated with *F* statistics in ANOVA contexts. If **D** is a matrix of Euclidean distances (*N* x *N*) based on a vector **Y** with *N* elements

for N individuals (thus $P = 1$ for a single gene), $\mathbf{G} = (YY')$ and Equation 3 reduces to Equation 1 since $rss_1 = \text{tr}[(\mathbf{I} - \mathbf{H})\ YY'(\mathbf{I} - \mathbf{H})]$ (McArdle and Anderson 2001; Wessel and Schork 2006).

Acknowledgements

The authors would like to thank Dr. Marti Anderson for advice and encouragement. NJS is supported in part by: The NHLBI Family Blood Pressure Program (FBPP; U01 HL064777-06); The NIA Longevity Consortium (U19 AG023122-01); the NIMH Consortium on the Genetics of Schizophrenia (COGS; 5 R01 HLMH065571-02); NIH R01s: HL074730-02 and HL070137-01; the UCSD Moores Cancer Center, and the Donald W. Reynolds Foundation (Helen Hobbs, Principal Investigator). This chapter appears as a reformatted version of the following submitted material to be published:

Ye C*, **Zapala MA***, Min Kang H, Wessel J, Eskin E, Schork NJ. High-Density QTL Mapping to Identify Loci Influencing Gene Expression Patterns in Entire Biochemical Pathways. *BMC Genomics*. 2007. Submitted.

Figure 17. B Cell Differentiation – Aggregate Association

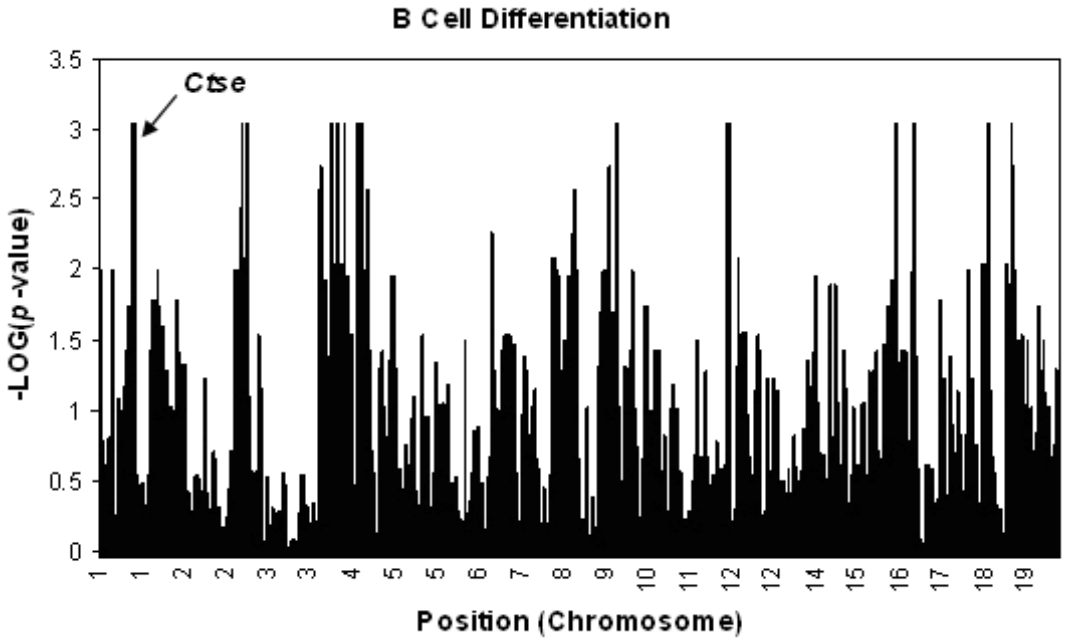


Figure 17. B Cell Differentiation – Aggregate Association. A whole genome aggregate association of gene expression signals in the B cell differentiation pathway. The x axis shows chromosomal position while the y axis shows the negative log of the aggregate *p*-value from permutation tests. The region harboring Cathepsin E (*Ctse*) is indicated above

Figure 18. Biosynthesis of Steroids – Multivariate Association

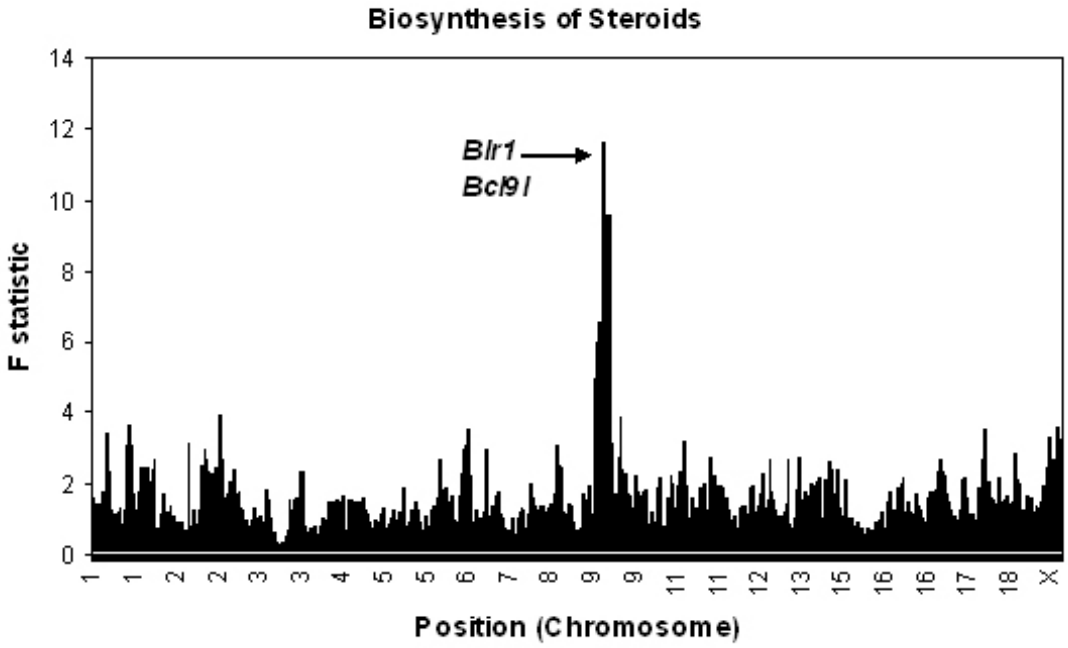


Figure 18. Biosynthesis of Steroids – Multivariate Association. A whole genome multivariate association of gene expression signals in the Biosynthesis of Steroids pathway. The x axis shows chromosomal position while the y axis shows the multivariate F-statistic. The largest F-statistic occurs over a haplotype block among the BXD recombinant inbred strains that contains both Burkitt lymphoma receptor 1 (*Blr1*) and B-cell CLL/lymphoma 9-like (*Bcl9l*).

Figure 19. SNP Marker rs6250833 - Aggregate Association

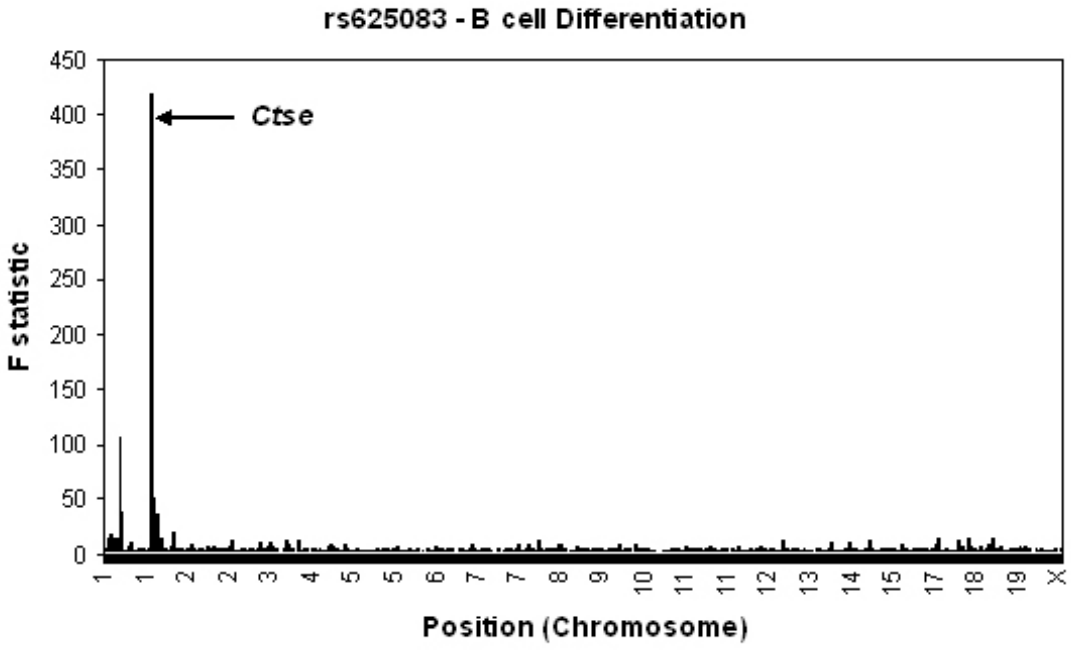


Figure 19. SNP Marker rs6250833 - Aggregate Association. This figure shows the genome-wide association of rs6250833, a SNP identified using aggregate statistics to be associated with the B cell differentiation pathway. The highest association is with *Ctse*.

Figure 20. EGFR in Cardiac Hypertrophy – Multivariate Association

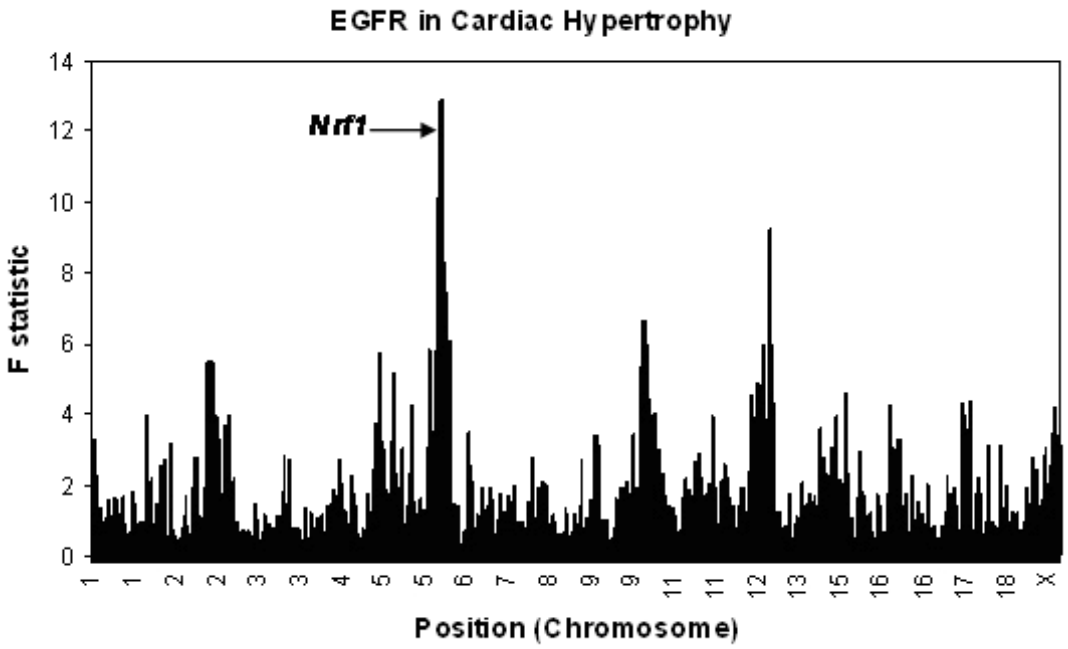


Figure 20. EGFR in Cardiac Hypertrophy – Multivariate Association. A whole genome multivariate association of gene expression signals in the Biocarta Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy pathway. The x axis shows chromosomal position while the y axis shows the multivariate F-statistic. The largest F-statistic occurs over a haplotype block among the BXD recombinant inbred strains that contains Nuclear Respiratory Factor 1 (*Nrf1*).

Table 8. Top 20 Univariate Associations

Top 20 SNP associations or gene-SNP pairs from the univariate ANOVA analysis. Bold rows indicate associations identified previously in Bystrykh, et al 2005. Multiple rsIDs are given when associations occur over stretches of linkage disequilibrium where each SNP will have the same *p*-value.

SNP	SNP Chr	SNP Position (Mb)	ProbeID	GeneID	Transcript Chr	Transcript Position (Mb)	P-value	Cis /Trans
rs13476148	**1**	**145**	**X104696_at**	**Ctse**	**1**	**133**	**4.89E-32**	**cis**
rs3669108, rs6375522, rs4136041	1	176	X93321_at	Ifi203	1	175	3.8009E-26	cis
rs6170159	2	144	X92839_f_at	Snrpb2	2	142	2.1339E-20	cis
rs13483308	**18**	**38**	**X92580_at**	**Hars**	**18**	**36**	**8.716E-20**	**cis**
rs3669108, rs6375522, rs4136041	1	176	X94224_s_at	Mnda	1	175	1.6119E-18	cis
rs3721918	14	63	X100929_at, X103799_at	Mtmr9	14	62	1.8177E-17	cis
rs3669108, rs6375522, rs4136041	1	176	X98465_f_at	Ifi204	1	175	7.601E-16	cis
rs3699123, rs13477251	**3**	**82**	**X101896_at**	**Cd1d2**	**3**	**87**	**1.354E-15**	**cis**
rs3663409, rs3151902	3	32	X93949_at	Gnb4	3	32	1.4713E-15	cis
rs6170159	2	144	X104225_at	Snx5	2	143	5.5896E-15	cis
rs3696278	**13**	**97**	**X95474_at**	**F2r**	**13**	**96**	**1.061E-14**	**cis**
rs6241531	1	9.99	X96156_at	AK048214	1	9.93	1.3599E-14	cis
rs13479414	**7**	**95**	**X96499_at**	**AI594671**	**11**	**39**	**2.134E-14**	**trans**
rs13480260	9	69	X95430_f_at	Spg21	9	65	2.6848E-14	cis
rs13478031	4	143	X95355_at	Sc5d	4	146	2.283E-13	cis
rs13482296	14	87	X95628_at	Diap3	14	85	6.1255E-13	cis
rs8254378	**9**	**123**	**X93397_at**	**Ccr2**	**9**	**123**	**7.584E-13**	**cis**
rs3693605	12	102	X104161_at	Cpsf2	12	102	1.7268E-12	cis
rs6388711, rs13480169	9	42	X102768_i_at	Sc5d	9	42	1.9148E-12	cis
rs13481399, rs13481403, rs3726096	12	39	X94937_at	Zfp277	12	40	2.0977E-12	cis

Table 9. Top 20 Associated Pathways from Aggregate Analysis

Top 20 pathways found by the aggregate statistic as sorted by the number of significant associations ($p < 0.0009$). Bold rows indicate pathways suspected to play a role in HSC differentiation.

Pathway	# Significant SNPs
mousepaths: MAP Kinase Signaling Pathway	27
BioCarta:IL 2 signaling pathway	26
GO:0045595:regulation of cell differentiation	**24**
GO:0006334:nucleosome assembly	23
mousepaths: Cell Cycle	**22**
GO:0008238:exopeptidase activity	21
GO:0042440:pigment metabolism	20
mousepaths: Dendritic _ Antigen Presenting Cell	19
BioCarta:Links between Pyk2 and Map Kinases	19
GO:0031497:chromatin assembly	19
BioCarta:TNF/Stress Related Signaling	18
GO:0030183:B cell differentiation	**17**
mousepaths: Cancer Drug Resistance _ Metabolism	17
GO:0046148:pigment biosynthesis	17
GO:0045333:cellular respiration	17
GO:0030120:vesicle coat	16
GO:0000226:microtubule cytoskeleton organization and biogenesis	16
mousepaths: Nitric Oxide	16
GO:0030097:hemopoiesis	**16**
GO:0051270:regulation of cell motility	16

Table 10. Top 10 Multivariate Associations

Top 10 SNP associations or pathway-SNP pairs from the multivariate matrix regression analysis and their corresponding pathway, bold rows indicate SNP associations not identified by the univariate analysis.

SNP	SNP Chromosome	SNP Position (Mb)	Pathway	Fstat	Pvalue	Univariate
rs8254378	9	123	BioCarta:Selective expression of chemokine receptors during T-cell polarization	33.761362	1.00E-16	Yes
rs13483308	18	38	KEGG:Aminoacyl-tRNA synthetases	20.558984	1.00E-16	Yes
rs13476148	1	145	mousepaths: Extracellular Matrix Adhesion Molecules	16.750239	1.00E-16	Yes
rs3721918	14	63	GO:0006470:protein amino acid dephosphorylation	15.77527	1.00E-16	Yes
rs6395893	**17**	**39**	**GO:0019884:antigen presentation, exogenous antigen**	**13.51825**	**1.00E-16**	**No**
rs3696278	13	97	GO:0007596:blood coagulation	13.396842	1.00E-16	Yes
rs3701429	**6**	**30**	**BioCarta:Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy**	**12.86895**	**1.00E-16**	**No**
rs3696264	9	44	KEGG:Biosynthesis of steroids	11.600127	1.00E-16	Yes
rs13481632	**12**	**106**	**BioCarta:EPO Signaling Pathway**	**10.82715**	**1.00E-16**	**No**
rs3669108	1	176	GO:0001649:osteoblast differentiation	10.795723	1.00E-16	Yes

References

Battle TE, Yen A. (2002). Ectopic expression of CXCR5/BLR1 accelerates retinoic acid- and vitamin D(3)-induced monocytic differentiation of U937 cells. *Exp Biol Med* 227:753-762.

Bryder D, Ramsfjell V, Dybedal I, Theilgaard-Monch K, Hogerkorp CM, Adolfsson J, Borge OJ, Jacobsen SEW. (2001). Self-renewal of multipotent long-term repopulating hematopoietic stem cells is negatively regulated by Fas and tumor necrosis factor receptor activation. *J Exp Med* 194:941.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC,Williams RW, Cooke MP, de Haan G. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* 37:225-232.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 2005, 37:233-242.

Cui X, Churchill GA. (2003). Statistical tests for differential expression cDNA microarray experiments. *Genome Biol* 4:210.

Gao Y, Camacho LH, Mehta K. Retinoic acid-induced CD38 antigen promotes leukemia cells attachment and interferon-gamma/interleukin-1beta-dependent apoptosis of endothelial cells: Implications in the etiology of retinoic acid syndrome. *Leuk Res*, in press.

Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. (2006). Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. *PloS Genetics*. 2:1182-1192.

Goeman JJ, Oosting J, Cleton-Jansen A, Anninga JK and van Houwelingen HC. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* 21:1950-1957.

Grubb SC, Churchill GA, Bogue MA. A collaborative database of inbred mouse strain characteristics. (2004). *Bioinformatics* 20:2857-2859.

Grupe A, Germer S, Usuka J, Aud D, Belknap JK, Klein RF, Ahluwalia MK, Higuchi R, Peltz G. (2001). In silico mapping of complex disease-related traits in mice. *Science* 292:1915-1918.

Hiçsönmez G, Tuncer MA, Güler E, Tan E, Tekeliolu M. (1993). The potential role of high-dose methylprednisolone on the maturation of leukemic cells in children with acute promyelocytic leukemia (APL). *Exp Hematol* 21:599-601.

Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol* 4:R70.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437:1299-1320.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.

Jansen RC, Nap JP. (2001). Genetical genomics: the added value from segregation. *Trends Genet* 17:388-391

McArdle BH, Anderson MJ. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82:290-297.

Morrison SJ, Qian D, Jerabek L, Thiel BA, Park IK, Ford PS, Kiel MJ, Schork NJ, Weissman IL, Clarke MF. (2002). A genetic determinant that specifically regulates the frequency of hematopoietic stem cells. *J Immunol* 168:635-642.

Ohl L, Henning G, Krautwald S, Lipp M, Hardtke S, Bernhardt G, Pabst O, Forster R. (2003). Cooperating mechanisms of CXCR5 and CCR7 in development and organization of secondary lymphoid organs. *J Exp Med* 2003, 197:1199-1204.

Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW, Lagler E, Korstanje R, Wang X, Nusskern D, Bogue MA, Mural RJ, Paigen B, Wiltshire T. (2004). Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol* 2:e393.

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* 2003, 422:297-302.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM,

Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 37:710-717.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-15550.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* 102:13544-13549.

Wade CM, Daly MJ. (2005). Genetic variation in laboratory mice. *Nat Genet* 37:1175–1180.

Weinberg JB, Mason SN, Wortham TS. (1992). Inhibition of tumor necrosis factor-alpha (TNF-alpha) and interleukin-1 beta (IL-1 beta) messenger RNA (mRNA) expression in HL-60 leukemia cells by pentoxifylline and dexamethasone: dissociation of acivicin-induced TNF-alpha and IL-1 beta mRNA expression from acivicin-induced monocytoid differentiation. *Blood* 79:3337-3343.

Wessel J, Schork NJ. (2006). Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis. *Am J Hum Genet* 79:792-806.

Ye C, Eskin E. (2007). Discovering tightly regulated and differentially expressed gene sets in whole genome expression data. *Bioinformatics* 23:e84-90.

Zapala MA, 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* 103:19430-19435.

CHAPTER 8

Summary and Conclusions

Summary

This thesis demonstrates the progress in utilizing gene expression microarray technology, whole genome SNP data and new statistical methods to analyze the brain from a genomic viewpoint. Through the integration of whole genome gene expression and SNP data with multivariate statistical techniques we can begin to understand the molecular complexity of the mammalian central nervous system. To appreciate the genomic architecture of the mammalian brain, we initially undertook a broad survey of the molecular gene expression patterns in the mouse brain to develop a genomic brain map. Once this genomic map had been established, a more in-depth analysis occurred that attempted to sort out subtleties in brain function and genomic transcriptional regulation. We analyzed gene expression in the brain across multiple individuals in multiple brain regions. We integrated this gene expression data with genome-scale sequence variation data in an expression QTL analysis to gain insights into genomic transcriptional regulation of various brain regions.

Performing these types of expression QTL analyses create numerous statistical and analytical issues. First, SNPs within a probe set region which might be in linkage disequilibrium with a close tagging SNP could affect expression results and lead to spurious associations. As eQTL studies become more prominent, methods that help correct for false associations due to sequence differences within the probe set will be increasingly important. We developed an algorithm, called GeSNP, and a user friendly website (http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi) to reliably detect sequence variations affecting probe hybridization. The algorithm was used to

improve eQTL results, identify sequence variations between inbred mouse strains, humans and chimpanzees, and cases and controls of inflammatory bowel disease.

Beyond the issue of sequence variation affecting array hybridization, there are two other potential drawbacks of eQTL studies. The first drawback concerns the number of statistical tests performed. Since thousands of gene expression signals are typically tested against thousands of SNP markers, there is tremendous potential for false-positives. The second drawback concerns the biological meaning of an association revealed by these studies. Previous eQTL approaches have mainly focused on treating each gene independently and ignoring the correlations between the expression patterns of genes involved in the same biochemical pathway or functional group. These two disadvantages are not unique to eQTL studies but are true for high-dimensional biological data analyses in general, such as gene expression data, DNA sequence data, proteomic data or functional magnetic resonance imaging (fMRI) data. In order to combat the pitfalls mentioned above, we developed a multivariate statistical technique that relates the distance between all pairs of individuals with respect to the high dimensional data of interest and tests linear hypotheses that consider whether additional factors collected on the individuals can explain variation in the distances. Beyond displaying the utility of this technique in both gene expression and eQTL analyses, we demonstrated that this technique has broad appeal and value by exploring the favorable statistical properties of this new test.

Mouse Molecular Brain Map

Our goal in this study was to understand how regional gene expression patterns in the brain are related to brain architecture and organization. We sought to identify relationships between brain regions based on both shared and restricted gene expression patterns. The gene expression data were analyzed using molecular classification algorithms, without pre-specified anatomical information, to define relationships between brain structures. To our surprise, we found that the gene expression patterns of the adult brain have a transcriptional "imprint" that is consistent with embryological origins and classic evolutionary relationships between subregions of the cortex. These results suggest that while the expression pattern for many genes may change dramatically during development, the brain retains a degree of gene expression patterning established during embryogenesis that is important for maintaining regional specificity and functional relationships between brain regions in the adult. Several studies of the developing brain have demonstrated that similar sets of genes are used to establish a particular anatomical region and to maintain the cell-cell relationships of the differentiated region. Thus, it may be that the roles of these genes in adulthood are similar to their roles during development. These roles include maintaining established phenotypes and connectivity of neuronal populations, as well as preserving barriers to the inappropriate migration of neurons from one region to another. We speculate that these genes continue to play an important role in the regional specification of functional units in the adult brain.

DNA variation and brain-region gene expression between inbred mouse strains

We investigated DNA polymorphisms and gene expression profiles of various brain regions in six inbred mouse strains. We noticed an enrichment of *cis*-acting transcriptional regulators among the strain-specific genes, while the brain region-specific genes seem to be mainly regulated by *trans*-acting elements. In addition, our data suggest that different inbred mouse strains have very different relative amounts of certain transcripts in some brain regions, indicating complex brain region-specific regulatory networks. Our findings shed light on regulatory mechanisms of gene expression in different brain tissues and strains on a genomic scale, and have important implications for the design and analysis of eQTL mapping studies. We have shown that the extent of global DNA sequence variation does not directly determine the extent of gene expression variation between inbred mouse strains.

Detecting Genetic Variation in Microarray Expression Data

There are numerous statistical problems that must be overcome in expression QTL analyses. One particular issue is the fact that sequence variations in the target sequence the array is probing can lead to artificial eQTL associations. In order to combat this issue and take advantage of existing gene expression data we created an algorithm and web program named "GeSNP" to computationally mine gene expression array data for sequence variations. We demonstrated that the GeSNP algorithm can identify sequence differences using array-based gene expression data. The approach is general to several Affymetrix gene expression array types and

applicable to the analysis of data obtained in different populations of genetically distinct individuals, including humans. The GeSNP program can be used not only to identify small sequence differences, such as single-base substitutions, but also larger deletions or insertions and genes with different splice forms. In addition to the identification of genetic variants, the analysis method described here may find its most immediate application in improving array performance and enabling arrays designed for one strain or species to be used more broadly. We have used this technique successfully in the past to improve the quality of gene expression data by masking probes that cover regions with potential sequence differences in both mouse and human studies. Moreover, GeSNP was used to minimize false positive eQTL associations in an inbred mouse study.

Multivariate Distance Matrix Regression in Gene Expression Analyses

Neurogenomic analyses are complex, as researchers often have multiple individuals, multiple brain regions and numerous gene expression signals. In order to analyze these multifaceted data sets, we developed a multivariate distance matrix regression (MDMR) technique for the multivariate analysis of gene expression data. The proposed analysis procedure can be used to emphasize the multivariate nature of the expression values of many genes in the same pathway or brain regions in the same neural circuit and treats the system being interrogated as a whole. It is novel from former gene expression analysis techniques in that it does not consider each individual gene in a univariate analysis which then analyzes the univariate results in aggregate. It

is arguable that physiological perturbations and variations are likely to "re-set" the coordinated expression patterns of many genes in order to reach biochemical or physiological homeostasis or equilibrium. Thus, the assessment of the similarity of global gene expression profiles of multiple samples with different features or exposures is appropriate.

Statistical properties of Multivariate Distance Matrix Regression

Despite the broad utility and appeal of this new MDMR technique in analyzing gene expression data, the statistical properties of MDMR, such as power and level accuracy, had not been explored. We performed extensive simulation studies which suggest that the MDMR analysis procedure has exceptional promise as an adjunct or alternative to standard multivariate analysis methods for use with modern high-throughput biological assays. The MDMR procedure is ideally suited for settings in which the number of variables collected on individual samples is much greater then the number of samples. Often, researchers are interested in analyzing multivariate data collected on a group of individuals as though that data were providing multivariate "profiles" of those individuals, rather than as data on a distinct set of variables requiring independent attention. Such settings are the rule, rather than the exception, in many modern biological experiments. Our studies show that the properties of the test statistics for pursuing MDMR analysis are quite good, in that they are well-behaved, exhibit an excellent level accuracy and have good power to detect a wide-range of multivariate phenomena. In addition, we confirmed that the *F*-statistic used to test

associations within the MDMR framework follows an *F*-distribution with an intuitive number of degrees-of-freedom which means that there is a computationally efficient alternative to permutation-based tests. This computational efficiency can be of great value if MDMR analyses are to be pursued in settings where repeated tests are to be performed, such as in testing associations between hundreds of thousands of DNA sequence variations and multivariate phenotypes in whole genome association studies.

Pathway oriented eQTL whole genome associations studies

MDMR analysis was pursued in a whole genome association study to incorporate biochemical pathway information into an expression QTL analysis. The results demonstrate the power of performing univariate, aggregate and multivariate statistical analyses together to interrogate expression QTL data. The univariate approach allows for the identification of individual genes whose expression appears genetically regulated. The aggregate approach facilitates the discovery of SNPs whose association is over-represented in a particular pathway. Finally, the multivariate analysis examines entire biological systems as they relate to DNA sequence variation. Thus, each analysis method offers unique and important information that could ultimately lead to new discoveries and hypotheses about genetic mechanisms involved in regulating gene expression as well as phenotypic expression. In this study we have focused on SNPs as the independent variables. However, the methods presented here can easily be extended to include phenotypic information of relevance. By associating pathways to both SNPs and phenotypes, researchers can begin to understand the

genetic basis of complex phenotypes whose expression is the result of multiple genes working in concert.

Conclusions

The studies described in this thesis focused on bringing to bear new genomic technologies and statistical techniques to the study of a complex organ, the brain. The mouse model was used as a surrogate to study the human brain and brought with it unique advantages, such as the ability to leverage the genetics of inbred mouse strains. These studies have advanced the field of neurogenomics which is dedicated to understanding the genetics of nervous system function and the regulation of complex behaviors. The mouse molecular brain map demonstrated that there were numerous embryonic patterning genes that maintained high levels of patterned expression into adulthood. Incorporating multiple individual gene expression patterns and integrating sequence variation data showed that gene expression varied more widely across brain region than across individuals leading to complicated brain region transcriptional networks that would invalidate whole brain eQTL studies. Proper statistical analysis is critical in these types of genomic studies. We developed an algorithm to identify sequence variation which can disrupt eQTL studies and explored a new multivariate technique for the analysis of complex gene expression patterns. This new multivariate technique was used to perform eQTL associations that included biochemical pathway information into the association. In the end, we developed novel analysis tools that facilitate neurogenomic studies, shed light on the genomic architecture and transcriptional regulatory machinery of the mammalian brain.