2001

Genetic Dissection of Behavioral and Neurogenomic Responses to Acute Ethanol

Aaron Wolen
Virginia Commonwealth University

School of Medicine
Virginia Commonwealth University

This is to certify that the dissertation prepared by Aaron R. Wolen entitled "Genetic Dissection of Behavioral and Neurogenomic Responses to Acute Ethanol" has been approved by his or her committee as satisfactory completion of the dissertation requirement for the degree of Doctor of Philosophy.

Michael F. Miles, M.D., Ph.D., Pharmacology and Toxicology

Aron H. Lichtman, Ph.D., Pharmacology and Toxicology

Mark A. Reimers, Ph.D., Biostatistics

Rita Shiang, Ph.D, Human and Molecular Genetics

Timothy P. York, Ph.D, Human and Molecular Genetics

Paul B. Fisher, Ph.D., Chair, Human and Molecular Genetics

Jerome F. Strauss, M.D., Ph.D., Dean, School of Medicine

F. Douglas Boudinot, Ph.D., Dean, School of Graduate Studies

February 1, 2012

GENETIC DISSECTION OF BEHAVIORAL AND NEUROGENOMIC RESPONSES TO ACUTE ETHANOL

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy at Virginia Commonwealth University.

by

AARON R. WOLEN

Bachelor of Science, University of Iowa, 2006

Director: Michael F. Miles, M.D., Ph.D.

Professor, Department of Pharmacology and Toxicology

Virginia Commonwealth University

Richmond, Virginia

February 2012

Acknowledgments

First and foremost I would like to express my sincere gratitude to my Ph.D. advisor, Dr. Michael Miles. I've learned so much from him, not only about science, but also what it means to be a scientist. And despite being one of the busiest people on the planet, Dr. Miles has always extremely generous with his time and his expertise. I feel very fortunate for having completed my graduate studies as a member of his lab.

I would also like to thank the members of my graduate committee who provided valuable suggestions that greatly improved the quality of my project. Discussions with Drs. York and Reimers substantially improved my understanding of the many statistical issues my project entailed. I would especially like to thank Dr. Shiang for all of her support since I started graduate school and her gentle reminders that various deadlines were approaching. I would also like to thank Dr. Michael Neale for supporting my work with his training grant from the National Institute of Mental Health.

My project has also greatly benefited from the expertise of our collaborators. I genuinely enjoyed working and corresponding with Dr. Michael Langston and Charles Phillips, who developed the paraclique algorithm. I am especially grateful to Dr. Rob

Williams for being so generous with his data and his time. Additionally, I'd like to thank the Linux gurus at the VCU Center for the Study of Biological Complexity, especially John Noble and Carlisle Childress, for all of their help with the high performance computer clusters used to perform many of the large scale analyses required for this project.

I would like to thank past and present members of the Miles lab for their friendship and all of the assistance they provided along the way; in particular, I'd like to thank Sean Farris, Tom Green, Jennifer Wolstenholme, Nathan Bruce, Paul Vorster, Megan O'Brien, JoLynne Harenza and especially Alex Putman, for handing down such an interesting project. I would also like to acknowledge George Walton, Andrew Beer, Roman Kotov, Ana Llopart and Josep Comeron, each of whom played a major role in my decision to pursue a career in biological research.

And finally, I'd like to thank Bernice Huang, who has been my constant companion throughout graduate school and made the entire journey more fun than it has any right to be; and my parents, Ralph and Betsy—it is only with their unconditional support and endless encouragement that any of this was possible.

Table of Contents

List of Figures

List of Tables

List of abbreviations

AvgDiff . . . . . .Average Difference

C. elegans . . . . . .*Caenorhabditis elegans*

E. coli . . . . . .*Escherichia coli*

Etanq1 . . . . . .ethanol-induced anxiolytic-like response QTL 1

trans-band . . . . . .*trans* eQTL-band

AA . . . . . .amino acid

AI . . . . . .advanced intercross

ANOVA . . . . . .analysis of variance

AUD . . . . . .alcohol use disorder

B6 . . . . . .C57BL6/J

BK . . . . . .big potassium, large conductance

BP . . . . . .biological process

BXD . . . . . .B6 × D2

CC . . . . . .cellular component

cDNA . . . . . .complementary DNA

Chr . . . . . .chromosome

CI . . . . . .confidence interval

CIE . . . . . .chronic intermittent ethanol exposure

CNS . . . . . .central nervous system

CPU . . . . . .central processing unit

CRE . . . . . .cAMP response element

CRF . . . . . .corticotropin-releasing factor

cRNA . . . . . .complementary RNA

Cy3 . . . . . .Cyanine 3

Cy5 . . . . . .Cyanine 5

D2 . . . . . .DBA2/J

DG . . . . . .dentate gyrus

DNA . . . . . .deoxyribonucleic acid

eQTL . . . . . .expression QTL

ErGeN . . . . . .ethanol responsive gene enriched network

F_2 . . . . . .second filial

FANS . . . . . .functional analysis of novel SNPs

FDR . . . . . .false discovery rate

FPM . . . . . .fat-pad mass

GABA . . . . . .γ-aminobutyric acid

GO . . . . . .gene ontology

GWA . . . . . .genome-wide association

HAP . . . . . .high-alcohol preference

HGNC . . . . . .HUGO Gene Nomenclature Committee

HSD . . . . . .honest significant difference

HUGO . . . . . .Human Genome Organization

IBD . . . . . .identical by descent

ILS . . . . . .inbred Long-Sleep

IP . . . . . .intraperitoneal

ISS . . . . . .inbred Short-Sleep

IVT . . . . . .*in vitro* transcription

KEGG . . . . . .Kyoto Encyclopedia of Genes and Genomes

LAP . . . . . .low-alcohol preference

LCMS . . . . . .likelihood-based causality model selection

LD . . . . . .linkage disequilibrium

LD box . . . . . .light-dark box

LOD . . . . . .logarithm of odds

LORR . . . . . .loss of righting reflex

LXS . . . . . .ILS × ISS

M430v2 . . . . . .Mouse Genome 430 2.0

MAS 4.0 . . . . . .Microarray Analysis Suite 4.0

MAS 5.0 . . . . . .Microarray analysis Suite 5.0

Mb . . . . . .megabase

MF . . . . . .molecular function

MIAME . . . . . .minimum information about a microarray experiment

MM . . . . . .mismatch

mRNA . . . . . .messenger RNA

NAc . . . . . .nucleus accumbens

ncRNA . . . . . .non-coding RNA

nt . . . . . .nucleotide

PCA . . . . . .principal component analysis

PCR . . . . . .polymerase chain reaction

PDT . . . . . .percent distance traveled in the light

PFC . . . . . .prefrontal cortex

PM . . . . . .perfect match

PTS . . . . . .percent time spent in the light

QTG . . . . . .quantitative trait gene

QTL . . . . . .quantitative trait locus

RI . . . . . .recombinant inbred

RMA . . . . . .robust multi-array average

RNA . . . . . .ribonucleic acid

RNA-seq . . . . . .high-throughput RNA sequencing

RNase . . . . . .ribonuclease

RQI . . . . . .RNA quality indices

S-score . . . . . .significance score

SAM . . . . . .Significance Analysis of Microarrays

SD . . . . . .standard deviation

SEM . . . . . .standard error of the mean

siRNA . . . . . .small interfering RNA

SNP . . . . . .single nucleotide polymorphism

TF . . . . . .transcription factor

TLA . . . . . .total locomotor activity

U74v2 . . . . . .Mouse Genome U74 2.0

UTR . . . . . .untranslated region

VMB . . . . . .ventral midbrain

VTA . . . . . .ventral tegmental area

WGCNA . . . . . .weighted correlation network analysis

List of genes

Crf-bp

CRF binding protein. 82

Drd1

dopamine receptor D1. 88

Drd2

dopamine receptor D2. 88

Dusp1

dual specificity phosphatase 1. 44

Egr2

early growth response protein 2. 44

Fos

FBJ osteosarcoma oncogene. 44

Gabrb2

GABA A receptor, subunit α2. 3, 55

Gapdh

glyceraldehyde-3-phosphate dehydrogenase. 39

Gria1

glutamate receptor, ionotropic, AMPA 1. xxix, 78, 84

Grm2

glutamate receptor, metabotropic 2. 90

Grm3

glutamate receptor, metabotropic 3. xxix, 79, 80, 90

Gsk3β

glycogen synthase kinase 3β. xxix, 56, 79, 121

Hsd11b1

hydroxysteroid 11-β dehydrogenase 1. 127

Hsp8

heat shock protein 8. 44

Jun

jun proto-oncogene. 44

Kcnma1

potassium large conductance calcium-activated channel,

subfamily M, α member 1. xxix, 55, 79, 90, 91

Kcnq5

potassium voltage-gated channel, KQT-like subfamily, member 5. 90, 91

Lcn2

lipocalin 2. 48, 50

Map4k5

mitogen-activated protein kinase 5. 115, 116, 118, 120

Mpdz

multiple PDZ domain protein. 6

Ncor1

nuclear receptor co-repressor 1. 78, 84

Nell2

NEL-like 2. xxix, 78, 84

Nin

ninein. 115, 116, 118–121

Npas4

neuronal PAS domain protein 4. 44, 56

Npy

neuropeptide Y. 3

Nr4a1

nuclear receptor subfamily 4, group A, member 1. 50

Nrg3

neuregulin 3. xxix, 90

Pygl

liver glycogen phosphorylase. 118–121

Rab3a

member RAS oncogene family 3a. 86

Scn1b

sodium channel, voltage-gated, type I β. xxix, 78, 82, 91

Sncb

synuclein β. xxix, 78, 82

Sos2

son of sevenless homolog 2. 115, 116, 118, 121

Trim9

tripartite motif containing 9. 115, 116, 118, 121

Abstract

GENETIC DISSECTION OF BEHAVIORAL AND NEUROGENOMIC RESPONSES TO ACUTE ETHANOL

By Aaron R. Wolen, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2012

Major Director: Michael F. Miles, M.D., Ph.D.

Professor, Department of Pharmacology and Toxicology

Individual differences in initial sensitivity to ethanol are strongly related to the heritable risk of alcoholism in humans. To elucidate key molecular networks that modulate ethanol sensitivity we performed a systems genetics analysis of ethanol-responsive gene expression in brain regions of the mesocorticolimbic reward circuit (prefrontal cortex, nucleus accumbens, and ventral midbrain) across the B6 × D2 (BXD) recombinant inbred (RI) panel, a highly diverse family of isogenic mouse strains before and after

treatment with ethanol.

Acute ethanol altered the expression of ≈2,750 genes in one or more regions and 400 transcripts were jointly modulated in all three. Ethanol-responsive gene networks were extracted with a powerful graph theoretical method that efficiently summarized ethanol's effects. These networks correlated with acute behavioral responses to ethanol and other drugs of abuse. As predicted, networks were heavily populated by genes controlling synaptic transmission and neuroplasticity.

Several of the most densely interconnected network hubs, including *Kcnma1* and *Gsk3β*, are known to influence behavioral or physiological responses to ethanol, validating our overall approach. Other major hub genes like *Grm3* and *Nrg3* represent novel targets of ethanol effects. Networks were under strong genetic control by variants that we mapped to a small number of chromosomal loci. Using a novel combination of genetic, bioinformatic and network-based approaches, we identified high priority *cis*-regulatory candidate genes, including *Scn1b*, *Gria1*, *Sncb* and *Nell2*.

The ethanol-responsive gene networks identified here represent a previously uncharacterized intermediate phenotype between deoxyribonucleic acid (DNA) variation and ethanol sensitivity in mice. Networks involved in synaptic transmission were strongly regulated by ethanol and could contribute to behavioral plasticity seen with chronic ethanol. Our novel finding that hub genes and a small number of loci exert major influence over the ethanol response of gene networks could have important implications for future studies regarding the mechanisms and treatment of alcohol use disorders.

Chapter 1

Introduction

1.1 Genetics of alcoholism

Alcohol use disorders (AUDs) are extremely prevalent; according to the most recent National Survey on Drug Use and Health (2011), an estimated 18 million Americans meet diagnostic criteria for an AUD. However, only a small subset of the wider population that regularly consumes alcohol will ever meet clinical criteria for alcohol abuse or alcoholism. While it is well established that AUD susceptibility is strongly influenced by genetic factors, which account for as much as 40–60% of the risk for developing an AUD (Heath et al., 1997; Kendler et al., 1994), family history is still the best predictor of an individual's risk for developing an AUD.

While population and family-based association studies have discovered a number of genetic markers linked to AUD susceptibility (Foroud et al., 2000; Hill et al., 2004; Reich et al., 1998), the highly complex and multifactorial nature of the disorder suggests that, independently, each of these associations accounts for only a small portion of the overall genetic variance. Moreover, the molecular mechanisms underlying the neuroplasticity accounting for AUD likely involves networks comprised of many more

genes than currently identified as affecting behavioral responses to ethanol in animal models or genetically associated with AUD in humans.

1.2 Traditional approaches to dissecting complex traits

The multiple genetic, environmental, and behavioral factors that play a role in the development of AUDs make it difficult to identify individual genes linked to these disorders. Still, as discussed below, some genetic risk factors associated with AUDs have been identified in genes that code for proteins involved in known biological pathways. Despite this progress, determining which genes may be the most relevant to developing therapeutic interventions for alcoholism has proven exceedingly difficult. The major obstacles being that gene/disease associations reveal very little about the underlying biology and any implicated gene variant explains only a tiny proportion of an individual's overall risk for AUD. Recent work focusing on the study of gene networks is helping to shed light on the molecular factors affecting complex diseases such as AUD.

1.2.1 Genome-wide association

The forward genetics approach proposed by Botstein et al. (1980) over 30 years ago has enabled the identification of specific gene variants that drive a variety of rare Mendelian disorders. However, the success of linkage analysis and positional cloning methods could not be duplicated when applied to more common diseases characterized by complex patterns of inheritance and lower levels of penetrance (Hirschhorn et al., 2001). genome-wide association (GWA) studies were proposed as an alternative approach to uncover common gene variants that underlie such complex disease (Risch and Merikangas, 1996). This approach involves identifying correlations between single nucleotide polymorphisms (SNPs) and a particular trait of interest across a population

for which genotypic and phenotypic information is available (Hirschhorn and Daly, 2005).

Hundreds of complex diseases and traits have been analyzed using this GWA design, which have produced many important links between genetic variants and human diseases (Altshuler et al., 2008). Notable successes include prostrate cancer (Haiman et al., 2007a,b), Crohn's disease (Barrett et al., 2008) and type 2 diabetes (Zeggini et al., 2008). GWA studies have also uncovered variants associated with susceptibility to AUDs in the *Gabrb2*, which provides the $\alpha 2$ subunit of γ-aminobutyric acid (GABA) A receptors (Dick et al., 2006; Edenberg et al., 2004), *Npy* receptors (Wetherill et al., 2008) and the classic ethanol metabolizing genes, alcohol dehydrogenase and aldehyde dehydrogenase (Cook et al., 2005; Kuo et al., 2008; Whitfield, 2002; Whitfield et al., 1998). Overall, however, the success of this approach has been mixed, and greater progress has been hindered by insufficient sample sizes, stratified populations, the involvement of rare gene variants that each contribute only small effects, and heterogenous phenotypic constructs.

1.2.2 Quantitative trait locus mapping

A similar forward genetics approach commonly used for studying animal models of complex traits is called quantitative trait locus (QTL) mapping. This approach involves measuring a particular quantitative trait across a genetically diverse population and scanning for associations between genotypic and phenotypic variation (Doerge, 2002; Grisel, 2000). Genomic regions that show a sufficiently strong association with a phenotype are considered QTL. The simplest, or most hopeful, interpretation of a mapped QTL is the implicated region harbors a single gene that affects the manifestation of the associated phenotype. However, it is quite possible that a QTL is actually driven

by multiple genes, non-coding RNA (ncRNA) species, epigenetic mechanisms, or a combination thereof.

While QTL mapping studies may be carried out in human populations, the inflated non-genetic variance contributed by each subject's unique environment and life experience introduces a tremendous amount of noise, making results difficult to interpret (Broman and Sen, 2009). For numerous reasons mice are a highly attractive model for the purpose of dissecting complex traits. Apart from their small size, relatively low maintenance costs and short gestation period (Peters et al., 2007), mice also provide an incredibly deep and ever-expanding arsenal of genetic tools.

As such, QTL mapping studies are typically conducted with animal models, and primarily inbred strains of mice and their various derivatives. For example, the C57BL6/J (B6) and DBA2/J (D2) inbred mice are frequently used in alcohol research because they clearly differ in various responses to ethanol, including development of functional tolerance (Grieve and Littleton, 1979), locomotor activation (Phillips et al., 1995), and sensitivity to withdrawal symptoms (Metten and Crabbe, 1994). Because the environmental conditions in these experiments can be controlled, phenotypic differences observed between the mouse strains can be largely attributed to genetic differences. QTL mapping studies then seek to detect the polymorphisms underlying the phenotype of interest by scanning for alleles that co-vary with the traits.

1.2.3 Recombinant inbred panels of mice

QTL can only be mapped in the presence of genetic variation, therefore QTL studies are often conducted using derivatives of inbred strains. A typical experimental design might involve characterizing a panel of second filial (F_2) progeny for a phenotype where the inbred parental strains differ significantly. QTL mapping could then commence after

genotyping each of the F_2 progeny. A special derivative of inbred strains, RI strains, are produced much the same as an F_2 panel but includes an additional phase of multiple generations of sibling inbreeding (Bailey, 1971). The result is a panel of novel inbred strains, each carrying a unique combination of the progenitor genomes.

Because these inbred animals are completely isogenic, each generation of progeny is a genetic clone of its forebears. As such, these genetic lines are essentially immortal, an incredible boon for scientific reproducibility, because experiments conducted in different laboratories can be directly compared. For this reason, the popularity of inbred mice has greatly encouraged inter-lab collaborations and more open data sharing practices. Furthermore, being inbred, each strain needs to be genotyped only once. In practice, this has meant that larger laboratories specializing in genotyping, such as the Wellcome Trust Sanger Institute and the Jackson Laboratory (see `http://cgd.jax.org/cgdsnpdb`), have genotyped many inbred strains or RI panels and made the results publicly available. This exemplifies the power of working with RI panels in QTL mapping studies; all acquired data is cumulative and directly relatable, regardless of where it originates.

The RI panels most widely used in alcohol research are the BXD and ILS × ISS (LXS) batteries of RI lines. The BXD lines, derived from the B6 and D2 inbred strains, currently have over 80 inbred strains (Peirce et al., 2004; Taylor, 1978; Taylor et al., 1999). The LXS strains were derived from the inbred Long-Sleep (ILS) and inbred Short-Sleep (ISS) inbred strains that were originally derived by selective breeding for sensitivity to ethanol sedation (DeFries et al., 1989; Williams et al., 2004). A large collection of behavioral, anatomical and neurochemical phenotypes derived from the BXD and LXS RI lines is maintained on GeneNetwork.

The molecular and genetic resources outlined above serve to greatly increase the power and resolution of QTL mapping for complex trait dissection. However, despite the wealth of genetic resources and the large number of QTL that have been identified,

the validation of corresponding QTL has greatly lagged behind (Flint et al., 2005). This difficulty largely stems from the lack of sufficient recombination events in existing mouse panels to reduce haplotype blocks, which can span up to several megabases (Mbs) and are generally much larger than what's observed in humans (Shifman et al., 2006). Additionally, small effect sizes complicate detection of a QTL as fine-mapping efforts proceed. The effect size issue may be due in part to the existence of multiple quantitative trait genes (QTGs) underlying QTLs detected by initial screens. Strategies such as derivation of congenic lines have been successful for fine mapping a number of ethanol traits and for identifying *Mpdz* as one of the first QTGs mapped for a mammalian behavioral phenotype (Buck et al., 1999; Fehr et al., 2002; Shirley et al., 2004). However, such approaches take large investments in time, animals and research expenditures. In many cases, even with derivation of congenic lines, the support interval may comprise several Mb and potentially hundreds of positional candidate genes. As described below, the use of whole-genome expression profiling has provided a powerful approach for mitigating some of the difficulties presented by traditional genetic QTL mapping approaches.

1.3 Genomic approaches to dissecting complex traits

Because of the technical obstacles impeding their more effective use, both GWA and QTL mapping studies to date have identified a deluge of disease-associated genetic loci but few causal genes. Moreover, even the most successful studies have failed to place the disease-associated genes in any kind of biological context that would serve to explain the underlying functional biology. Without elucidating the complex interactions of the molecular phenotypes that stand between genetic variation and disease, it will be difficult or impossible to develop new and effective approaches to treating such diseases.

The emerging field of systems biology is tackling this immense challenge by studying networks of genes, proteins (Rual et al., 2005), metabolites (Nielsen and Oliver, 2005) and other molecular phenotypes that represent models of genuine biological pathways. Studying complex diseases in terms of gene networks rather than individual genes or genomic loci should aid in dissecting complex diseases by identifying molecular pathways that are perturbed by genetic variation and ultimately mediate the disease-associated phenotypes (Schadt, 2009). Furthermore, networks found to mediate relevant endophenotypes in animals models should be highly generalize to humans, as several studies have indicated the gene networks are evolutionarily conserved (Miller et al., 2010; Oldham et al., 2008; Stuart et al., 2003).

1.3.1 Defining disease using high-throughput molecular profiles

Platforms for high-throughput approaches for all these types of molecular profiling have become increasingly commonplace. Concurrently, methods for analyzing data produced by these technologies are constantly evolving, yielding results that are simultaneously more sensitive and more specific. As a result, researchers are better able to appreciate systems-level changes associated with disease. Of these various high-throughput profiling techniques, microarray-based gene expression platforms have featured most prominently in biomedical research to date. Through an unbiased profiling of the transcriptome, microarray expression studies allow researchers to identify patterns of gene expression associated with a disease.

In some cases, such patterns can better define a complex phenotype by identifying disease subtypes. For example, microarray analysis of breast cancer tumors identified gene expression signatures that predict patient prognosis and therefore help physicians tailor treatment regiments (van 't Veer et al., 2002). From a basic research perspective,

microarray expression profiles can help tease apart the complex interactions that underlie the development of a disease by implicating a subset of genes whose regulation is altered with the disease. With this information, it may become feasible to reconstruct the underlying biological pathways and enhance understanding of disease etiology.

Genomic approaches have been applied directly to alcoholism by studying postmortem human brain tissue isolated from alcoholics and matched control subjects using gene expression microarrays. This has revealed novel information about changes in the brain's transcriptome that are associated with chronic ethanol consumption. One of the findings was a significant deregulation of genes encoding proteins that synthesize and maintain myelin (Lewohl et al., 2000; Mayfield et al., 2002). However, the nature of these studies makes it impossible to determine whether such gene expression deviations actually are risk factors that contribute to AUDs or simply represent molecular consequences of excessive alcohol consumption that are unrelated to the behaviors constituting alcoholism.

1.3.2 Genomic analyses of AUD models

Animal models can greatly assist in this analysis by allowing for experiments that are far more informative and, consequently, too invasive to be performed with humans. Although animal models could never replicate a phenotype as complex as alcoholism, they can mimic certain facets of the trait (Bennett et al., 2006), which can then be associated with specific expression signatures using gene expression microarrays. For example, a genetic predisposition for alcoholism may entail a stronger than average preference for alcoholic beverages. This particular facet of alcoholism is captured by rodent models that were selectively bred to maximize a penchant or an aversion to ethanol, such as the aptly name high-alcohol preference (HAP) and low-alcohol

preference (LAP) mice (Grahame et al., 1999). In order to identify genes that may alter the perceived desirability of ethanol, gene expression microarrays were used to compare the brain transcriptomes of HAP and LAP mice, along with several other inbred mouse strains that drastically differ in ethanol preference (Mulligan et al., 2006). This important study identified a diverse array of molecular pathways associated with differences in ethanol preference. Some of the genes that had the largest effect size were related to neuronal function and to cellular homeostasis.

Another important facet of a genetic predisposition to alcoholism is a comparatively blunted sensitivity to the effects of ethanol. As studies have shown that individuals who are initially less sensitive to acute ethanol are more likely to have a family history of alcoholism and are at greater risk for developing an AUD (Schuckit, 1984, 1994). As mentioned earlier, the B6 and D2 inbred mice are frequently used in genetic studies of ethanol sensitivity. For this reason, Kerns et al. (2005) used microarray expression studies to dissect the effect of acute ethanol on the brain's transcriptome using the B6 and D2 inbred mouse strains. The investigators analyzed three brain regions involved in the mesocorticolimbic reward pathway: prefrontal cortex (PFC), nucleus accumbens (NAc) and ventral midbrain (VMB). For each region analyzed, the study identified a specific set of gene modules whose expression was altered in response to acute ethanol exposure. These gene modules were significantly enriched for genes involved several retinoic acid signaling, neuropeptide expression and glucocorticoid signaling. Moreover, similar to the microarray studies of postmortem human alcoholic brains (Lewohl et al., 2000; Mayfield et al., 2002), several genes involved in myelination were robustly altered by alcohol exposure, particularly in the PFC (Kerns et al., 2005).

In examining the responses to acute or chronic alcohol exposure in rodent brains, these and numerous other genomic studies have enhanced the understanding of the *ethanol transcriptome* and provided a more comprehensive picture of the genes and

molecular pathways that contribute to specific facets of AUD than what is possible with studies of postmortem human brains (Daniels and Buck, 2002; Mulligan et al., 2011; Rimondini et al., 2002; Saito et al., 2004; Treadwell and Singh, 2004). Moreover, they have effectively demonstrated how gene expression microarrays can help narrow the information gap that exists between DNA variation and complex diseases. However, prioritizing the long lists of genes produced by comparative microarray studies conducted in either species has proven exceedingly difficult. Given the high costs associated with performing molecular validation experiments, an effective strategy for prioritizing candidate genes is crucial. Investigators therefore have used more systems-level approaches that combine genetic, genomic, and pharmacological methods to better delineate gene networks related to ethanol behavioral phenotypes.

1.4 Gene network analysis

The previous section mentioned several studies that used gene-expression microarrays to define lists of genes responding to ethanol or otherwise relevant to AUDs. Although these studies have provided important biological insights, the question of how such lists can be used to further advance understanding of a complex disease is not easily answered. Network-based approaches can greatly improve the interpretability of differential gene-expression results by providing information about the relationships between genes.

Networks are systems of interconnected components. For example, the World Wide Web is a global network of computers sharing documents connected by hyperlinks; road maps are visualizations of city networks connected by highways; social networks are groups of people connected through friendships; cellular signaling pathways are groups of proteins connected through molecular interactions (Junker and Schreiber, 2008). Placing such complex systems within a network framework makes it possible to formally

analyze the relationships that constitute these systems. Gene networks typically are visualized as mathematical graphs—that is, a collection of vertices and edges, where genes are represented by nodes and the lines connecting the nodes indicate that some relationship exists between the genes.

Many published network analyses of gene groups use information about pre-existing biological relationships, which may be derived from sources such as literature co-citation analysis (Rajagopalan and Agarwal, 2005), protein-protein interaction databases (Rual et al., 2005), or gene ontology (GO) groupings (Ashburner et al., 2000). Some commercial tools are available for such studies, such as Ingenuity Pathway Analysis (Ingenuity Systems, Redwood City, CA). Two recent human association studies effectively demonstrated the potential benefits of incorporating such information, by modifying the typical GWA strategy and only scanning for associations within groups of functionally related genes, rather than genome-wide. The first of these studies discovered that cognitive ability, a complex phenotype with a large genetic component, was significantly linked to genes encoding heterotrimeric G-proteins (Ruano et al., 2010). The second study found that genes related to glutamate and GABA signaling collectively contributes to alcohol dependence (Reimers et al., 2011).

However, although such approaches provide categories for interpreting the genomic data, they also force such interpretation into the mold of pre-existing information, potentially limiting the unbiased nature of genomic studies. Genomic data collected with high-throughput molecular profiling presents the opportunity to derive novel gene-gene interactions. The maturity of gene expression microarrays relative to similar technologies designed to measure other molecular phenotypes on a genomic scale has meant that gene networks are primarily rendered as gene co-expression networks. In the context of gene co-expression networks, links between nodes typically indicate that the expression levels for two genes are strongly correlated with one another across whatever

conditions an experiment entails (e.g. across tissues, time points, treatments, individuals, etc.). Each link in a gene network essentially represents a testable hypothesis that can be validated through follow-up molecular experiments. And indeed, co-expression networks have been used to identify protein interactions that are novel (Scott et al., 2005) and conserved across species (Stuart et al., 2003).

Various novel and innovative methods exist for generating gene co-expression networks. In their simplest form, however, gene co-expression networks can be constructed by calculating Pearson correlations between all gene pairs and applying a cut-off threshold to determine which genes should be connected. The simplicity of this approach makes it an appealing choice for conducting a first round of analyses. Section 3.1 provides a more detailed description of gene network construction methods.

A valuable advantage of such network-based approaches is that the relative importance of specific genes can be assessed in part, by the context of their surrounding interactions. A variety of calculations can be used to gauge the importance of nodes to the network as a whole (Dong and Horvath, 2007). The simplest measurement is determined by the degree of connectivity—that is, the number of other genes the node is connected to in the network. However, a gene's "position" in the network also is an important consideration. For example, a gene that served as the sole connection between two otherwise independent gene networks would rank fairly low on a priority scale based on connectivity alone, despite being an important channel of inter-module communication. A measurement of betweenness centrality (Girvan and Newman, 2002) can highlight such a gene by determining the frequency with which a node is included in the shortest paths between all possible node combinations. There is a growing body of evidence suggesting that hub genes are of particular importance to genetic networks. For example, introducing null mutations into hub genes negatively impacted the hardiness of *Escherichia coli* (*E. coli*) to a much greater extent than did mutations of randomly

selected genes (Cooper et al., 2006). This may be explained by an observation made in *Caenorhabditis elegans* (*C. elegans*), showing that hub genes participated in a variety of canonical signaling pathways (Lehner et al., 2006). In a genetic network study of mouse liver, hypothalamus and adipose tissue, hub genes were also found to be highly connected nodes across all three expression datasets (Dobrin et al., 2009).

1.5 Genetic dissections of genomic data

1.5.1 Molecular QTL

Another important early advancement toward a more systems-level approach to identifying disease-associated genes was the application of gene mapping methods to high-throughput molecular data, making it possible identify causal links between molecular phenotypes and genomic regions. Like classical physiological or behavioral phenotypes, genetic factors influencing high-throughput measures of transcript, protein and metabolite abundance can be identified by QTL mapping. The promise of this approach was first demonstrated by a study of *Zea mays* (maize) proteins (Damerval et al., 1994), in which 2-dimensional polyacrylamide gels were used to separate 72 proteins and measure their relative abundance levels across a population of 60 F_2 individuals. Looking for associations between these measurements and a panel of 100 genetic markers, Damerval et al. identified QTLs significantly influencing the abundance of over half of the analyzed proteins. Furthermore, this study effectively demonstrated the potential of this approach to provide unprecedented insight into structural complexities of quantitative trait regulation, by determining not only the number of QTL influencing a given trait, but also characterizing the dominance effects and epistatic interactions between QTL, and uncovering genetic regulators driving the co-expression between proteins with

highly similar expression patterns.

Analyzing the genetic regulation of such molecular phenotypes offers a much closer look at the biological processes that drive the variation in quantitative traits. By extending these analyses to include high-throughput molecular phenotypes, such as those generated by gene expression microarrays, it becomes possible to map out entire molecular networks or signalling pathways that underlie complex traits. The strategy of performing genetic linkage analysis on genome-wide molecular profiles was formalized and termed *genetical genomics* by Jansen and Nap (2001). This proposal primarily focused on gene-expression microarrays and posited that mapping expression QTLs (eQTLs) would enable researchers to construct robust gene networks as well as link these networks to metabolic or other phenotypes. The investigators also suggested that eQTL mapping could greatly aid in the identification of candidate genes underlying classical QTLs for disease traits.

The first study to carry out QTL analysis across gene expression microarray profiles was published using an experimental cross between two strains of *Saccharomyces cerevisiae* (Brem et al., 2002). Results from this landmark paper shed a great deal of light on the genetics basis of gene expression in a complex organism. Of the 6,215 genes measured, 1,528 were differentially expressed between the progenitor strains and 570 showed significant linkage to at least one locus. Importantly, the authors noted that the power to detect eQTLs for a gene is a direct function of the number of loci regulating that gene and the relative contribution made by each locus. By comparing results derived from empirical computer simulations to their observed data, they determined the majority of differentially expressed genes were likely regulated by at least 5 eQTL. Subsequently, several investigations applied the approach to mammalian systems (Schadt et al., 2003; York et al., 2005), including brain gene expression (Chesler et al., 2005, 2003).

1.5.2 *cis* and *trans* eQTL

These early genetical genomics studies also characterized the two major classes of eQTLs, labeled *cis* and *trans* eQTLs, which differ with respect to the position of the eQTLs relative to the gene whose expression is altered. A *cis* eQTL is located at the same site of the genome as the gene under study. In contrast, a *trans* eQTL can be located elsewhere in the genome, away from the gene whose expression is altered. A prototypical example of how a *trans* eQTL could manifest involves TFs: a SNP at the DNA-binding domain of a TF can affect the TF's ability to recognize and bind its recognition sites, causing altered expression of all genes regulated by this TF (Figure 1.1). In other words, the abundance of all transcripts from those genes would co-vary with the TF SNP. Such a case might be recognized by a clustering of *trans* eQTLs at the site of the causal polymorphism, sometimes referred to as a *regulatory hotspot* or *trans* eQTL-band (*trans*-band). The identification of *trans* eQTL clusters can be a powerful approach for identifying key regulators underlying a complex trait of interest.

The genes comprising *trans* eQTL clusters often have biological functions that have been conserved among species, suggesting that these *trans*-bands may have a biological relevance. Accordingly, the search for *trans* eQTLs may allow researchers to identify biological functions associated with complex traits through defining the gene networks that comprise a *trans*-band. For example, Mozhui et al. (2008) have dissected a *trans* eQTL cluster on distal mouse Chr 1 and identified a candidate gene that they propose has a major influence on the expression of linked gene networks and a diverse group of neurobiological phenotypes with QTLs located in the same region, including several related to ethanol and other drugs of abuse.

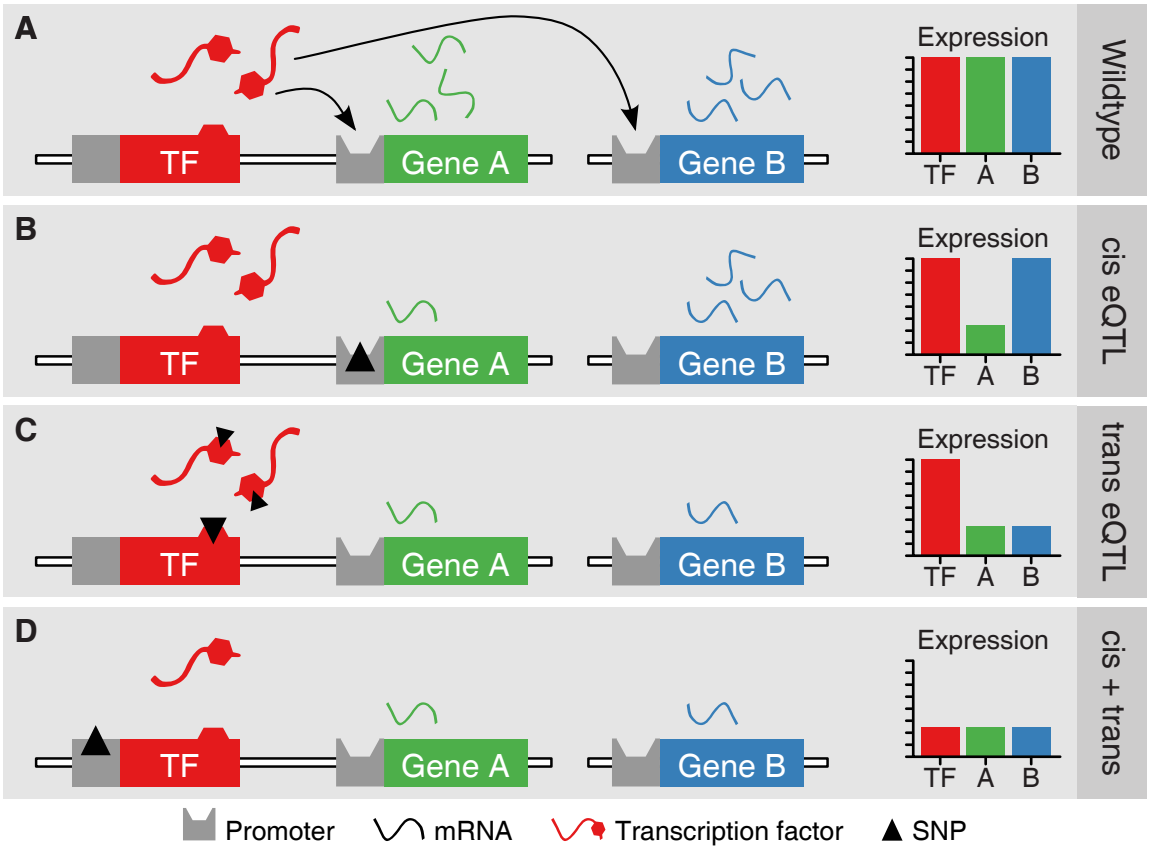

Figure 1.1. *cis* versus *trans* eQTL diagram. The left-most gene (red) codes for a TF that activates the transcription of genes A (green) and B (blue). **A.** In the wildtype scenario all genes are expressed at their full potential, as indicated by the bar graph on the right. **B.** A SNP within gene A's promoter hinders TF binding, causing a reduction in the rate at which it is expressed, while gene B is unaffected. Thus, variation in the expression of gene A is associated with a *cis* eQTL through the actions of a local *cis*-acting polymorphism. **C.** A SNP within the TF gene's DNA binding region (hexagon), hinders binding with all downstream promoters, regardless of whether the regulated gene is located near the TF gene, like gene A, or located elsewhere in the genome, like gene B. In fact, all genes regulated by this TF would potentially be linked to a *trans* eQTL at the site of this TF polymorphism. However, note that gene A's proximity to the TF gene would make it difficult accurately classify its eQTL as *cis* or *trans*; follow-up molecular experiments would be necessary to distinguish between the two possibilities. **D.** A SNP within the TF gene's promoter would manifest as a *cis* eQTL for the TF gene itself and a *trans* eQTL for all genes regulated by this TF downstream.

1.5.3 Integrative strategies for identifying genetic risk factors

The integration of expression eQTL and classical QTL data enables identification of key markers of disease-causing variants. The effectiveness of this approach was demonstrated by a genetical genomics analysis of liver expression data from a population of mice placed on a high-fat diet (Schadt et al., 2003). The purpose of this was diet was to model an obesity-like phenotype, which was measured using fat-pad mass (FPM). QTL mapping for FPM revealed a significant QTL on Chr 2 that also harbored over 400 eQTLs. By scanning this region for *cis* eQTL-linked genes that were also strongly correlated with FPM, the researchers were able to identify two novel obesity candidate genes.

Saba et al. (2006) used a similar approach to identify candidate genes for alcohol preference and acute functional tolerance to alcohol. This large-scale study included several selectively bred mice to maximize divergence in ethanol phenotypes, including the HAP and LAP inbred mice, as well a subset of the BXD RI family. Applying microarray expression profiles using messenger RNA (mRNA) samples obtained from the entire brain, the investigators identified independent lists of genes whose expression differed between the HAP and LAP strains and between the BXD strains with high and low levels of acute functional tolerance. Expression QTL mapping was then conducted for these differentially expressed genes using the BXD expression and genotypic data. They identified high-priority candidate genes by screening for differentially expressed genes with *cis* eQTL that overlapped previously mapped behavioral QTLs for either alcohol preference (Belknap and Atkins, 2001) or acute functional tolerance (Kirstein et al., 2002).

The rationale for prioritizing candidate QTGs on the basis of their having *cis* eQTLs located at the same sites as classical QTLs is based on the hypothesis that the variability of a complex phenotype is linked to a particular locus because the causal gene is being

produced in variable quantities through a *cis*-acting polymorphism. There is increasing evidence that supports the importance of gene-expression variability in regulating complex traits and interpreting associations between polymorphisms and complex traits (Emilsson et al., 2008; Kathiresan et al., 2008; Schadt et al., 2008). In fact, Recent evidence indicates that SNPs associated with a variety of complex traits are more likely to contain *cis* eQTLs than would be expected by chance alone (Nicolae et al., 2010). This indicates that the importance of expression variability in complex trait regulation is not limited to genetic model systems and that it may be possible for GWA and QTL mapping studies to improve their track record by incorporating expression data.

1.6 Summary

We have discussed how traditional QTL mapping and GWA studies can benefit from systems-biological approaches by filling in critical information about the molecular phenotypes that stand between DNA variation and complex disease. Incorporating data from high-throughput molecular profiling technologies, like gene expression microarrays, can better define a disease by identifying groups of genes that respond to or co-vary with disease-associated traits. Network analysis of disease-associated genes allows us to move beyond dichotomous gene lists, partially reconstruct the underlying molecular pathways and prioritize genes based on their importance to the larger network. Applying QTL mapping to each gene's expression trait, then makes it possible to identify the genomic regions that regulate each gene's expression and uncover the existence of regulatory hotspots that exert enormous influence over a gene network.

The project described in this thesis utilized the integrative genomic strategies described above to better define the mesocorticolimbic reward pathway's transcriptional response to acute ethanol. These efforts are a direct extensions of the work published

by Kerns et al., which provided the initial characterization of acute ethanol's impact on gene expression in the PFC, NAc and VMB. Here, we have used microarray expression data for the same three brain regions obtained for a large subset of the BXD RI panel. In Chapter 2 this data-set is used to identify the genes whose pattern of expression across the BXD panel is robustly altered by ethanol exposure. Chapter 3 characterizes a set of tightly coordinated gene co-expression networks that largely constitute the transcriptional response of the PFC to ethanol and identifies a small number of genetic loci that represent the key regulators of these networks. Chapter 4 describes the fine-mapping of a QTL for the anxiolytic-like response to acute ethanol and the use integrative genomic approaches to identify a single high priority candidate QTG.

1.7 Resources

Much of this project relied upon resources generated by previous members of the Miles laboratory and other collaborators. Dr. Alex Putman began the genetic analysis of the anxiolytic-like response to acute ethanol as part of his PhD project (Putman, 2008). The B6, D2 and BXD RI strains he assayed for this project also provided the brain tissue samples used to generate the PFC, NAc and VMB microarray expression datasets that are used throughout this thesis. The actual generation of the microarray expression data was expertly performed by Paul Vorster and Nathan Bruce. The BXD genotype data was generated by Williams et al. (2001) and Shifman et al. (2006) and is publicly accessible from GeneNetwork at `http://genenetwork.org/genotypes/BXD.geno`. Sequence data used in the B6/D2 SNP analyses was generated by Dr. Robert Williams and kindly provided by Dr. Xusheng Wang from his laboratory. This data can now be queried on GeneNetwork using the SNP browser. Additionally, Dr. Rob Williams provided the database of BXD phenotypes from GeneNetwork, used in the analysis

described in section 3.6.

The paraclique networks described in Chapter 3 were constructed in collaboration with Dr. Michael Langston (University of Tennessee), who developed the novel approach, and his graduate student, Charles Phillips. A web-based implementation of this software is available at `http://grappa.eecs.utk.edu/`.

All other analyses presented here were conducted using the open source R environment for statistical computing (R Development Core Team, 2011) and the genomic data analysis tools for R developed by the Bioconductor project (Gentleman et al., 2004). With the exception of network figures, which were rendered in Cytoscape (Shannon et al., 2003), all visualizations were generated in R using ggplot2 (Wickham, 2009). In most cases, ColorBrewer palettes were used for qualitative scales（www.colorbrewer.org).

Chapter 2

Neurogenomic response to acute ethanol

2.1 Microarray technology

Much of this thesis is devoted to the analysis of microarray gene expression data. These analyses took a variety of forms and included assessing microarray quality, generating expression summaries, integrating array data from disparate platforms, performing expression QTL mapping and extracting gene co-expression networks. As such, a brief review of the technology will greatly improve understanding of subsequent sections.

The term 'microarray' generically describes a technology that allows for biochemical assays to be performed in a massively parallel fashion. The array itself is typically a glass or plastic substrate stippled with a grid of molecular probes designed to specifically bind with complementary targets. The targets are often linked to chemiluminescent markers, making it possible to record successful probe/target hybridization events by scanning the array and analyzing the digital image using spot finding algorithms. Because the identity of every probe on the array is known, the fluorescence intensity of each spot can be used to extract quantitative information about the corresponding molecular target. Through the application of this microarray basic strategy, researchers have been able

to obtain measurements of DNA, mRNA, splicing events, TF binding, DNA methylation and other molecular events on scales not previously possible.

2.1.1 cDNA spotted microarrays

Although microarrays can be utilized to perform a variety of high throughput molecular assays, they are predominantly used in genetic investigations and specifically for the measurement of gene expression, providing an unbiased snapshot of intracellular mRNA transcript levels. The first modern DNA microarray study measured the expression of 45 genes in *Arabidopsis thaliana* (Schena et al., 1995). This study was conducted using spotted complementary DNA (cDNA) microarrays, one of the two principle gene expression microarray technologies. Probes for spotted cDNA microarrays are generated by polymerase chain reaction (PCR) amplification of cDNA libraries and printed on the array using automated robotic arms (Lennon and Lehrach, 1991). Although commercial cDNA libraries are available for producing spotted cDNA microarray probes, individual investigators may supply their own custom libraries, yielding microarrays completely tailored to address a specific research question. Studies typically utilize cDNA microarrays to determine the relative expression of genes between experimental conditions. This is achieved by labeling the samples with different fluorophors, most commonly Cyanine 3 (Cy3), which fluoresces green, or Cyanine 5 (Cy5), which fluoresces red. The two samples are then mixed and washed over a single microarray, where they competitively hybridize to the spotted probes. The intensity of each fluorescent signal is measured using laser-scanning microscopes and then compared, making it possible to determine whether any measured transcript is up- or down-regulated across experimental conditions (Shalon et al., 1996). While cDNA spotted microarrays broke new ground in the area of high-throughput gene expresion profiling, most recent work has utilized

alternative microarray platforms that rely on synthetic oligonucleotides affixed to beads (Kuhn et al., 2004) or glass slides (Lockhart et al., 1996).

2.1.2 Synthetic oligonucleotide arrays

Shortly after Schena et al. (1995) published their study of *A. thaliana* gene expression using cDNA spotted microarrays, Lockhart et al. (1996), from the commercial company Affymetrix (Santa Clara, CA), proposed an alternative DNA microarray technology for measuring gene expression using an *in situ* synthesis approach that essentially 'grows' oligonucleotide probes directly on the microarray substrate. This is achieved through a photolithographic process, which begins with a dense grid of synthetic linkers affixed to the substrate. These linkers terminate with a 5′-hydroxyl group that is initially blocked by a photo-senstive cap that is removed upon exposure to light (Pirrung et al., 1998). Any nucleotides washed over the array will chemically bond to linkers whose protective caps have been removed. The incorporated nucleotides are also modified with a photo-sensitive cap, ensuring no further synthesis will occur until a probe has undergone a subsequent round of light exposure. After repeating these steps for the remaining bases, every probe will have incorporated a single nucleotide (nt). By using a sequence of specially perforated masks that selectively expose probes to the light source, oligonucleotide synthesis can proceed simultaneously for all probes, one nucleotide at a time, requiring $4 \times N$ cycles to construct probes that are N nucleotides long (Lipshutz et al., 1999). Since synthetic oligonucleotide probes are, in fact, synthetic, their construction conveniently bypasses the need for cDNA libraries or to generate large quantities of purified PCR products, which are inherent to spotted cDNA arrays. Instead, synthetic oligonucleotide probes are designed *in silico* using available sequence data. Oligonucleotide microarrays further differ with cDNA spotted microarrays in that

only a single labeled sample is hybridized to each microarray. Relative expression is therefore determined using a series of spike-in controls, internal standards and through statistical comparisons with other microarrays.

2.1.3 The Affymetrix GeneChip system

Synthetic oligonucleotide arrays are commercially available from Affymetrix for a variety of purposes. The GeneChip® system represents Affymetrix's line of oligonucleotide arrays intended for measuring mRNA transcript abundance. This thesis is primarily concerned with gene expression data generated using the Mouse Genome 430 2.0 (M430v2) GeneChip, which provides expression measurements for $> 39,000$ transcripts across the mouse genome. Each of these transcripts is targeted by at least one set of oligonucleotide probes that are complementary to a different 25-base region of the same transcript.[1] The use of multiple probes per transcript is a valuable aspect of the GeneChip system, as collectively they can provide a more reliable measurement of transcript abundance that is more robust against individual probe aberrations. For example, if a single probe fell within a region contaminated by a bizarre spatial artifact, skewing its intensity level, the rest of the probe-set would be unaffected, since probes within a set are randomly distributed across the array.

In addition to the probes designed to be perfect complements of a particular transcript, probe-sets also include a second set of probes that are identical to the perfect match (PM) probes, with the exception of the 13^{th} nucleotide, which is intentionally swapped to disrupt the perfect complementarity. The intention of including mismatch (MM) probes was to provide a measurement of non-specific binding. However, as discussed below, MM probes capture signal as well as non-specific binding, and occasionally

[1] In practice, *most* probes within a probe-set target different regions of a transcript. However, it is not uncommon to come across probes with overlapping target sequences. In some cases the the overlap is extensive.

register higher intensity levels than their PM counterparts (Irizarry et al., 2003).

2.2 Gene expression microarray analysis

As mentioned earlier, microarray is an umbrella term that accurately describes a wide variety of technological platforms. Similarly, *microarray analysis* could be referring to an equally diverse set of topics. However, any subsequent mention of *microarrays* in this text is referring specifically to the Affymetrix GeneChip system unless stated otherwise. Even if limited to the discussion of Affymetrix microarrays, the analysis of gene expression microarray data is an incredibly broad and ever-evolving topic that is far beyond the scope of this thesis. Fortunately, many excellent reviews of microarray analysis strategies have been published that provide a much more thorough overview than could be accomplished here, many were written by leaders in the field, including Brown and Botstein (1999), Miles (2001), Quackenbush (2001), Churchill (2002) and Reimers (2010).

2.2.1 Sample preparation

The process of obtaining gene expression data using Affymetrix GeneChips begins with purified total ribonucleic acid (RNA) samples. The mRNA molecules are reverse transcribed into cDNA using PCR primers comprising the T7 bacteriophage promoter and a string of deoxythymidines. The samples are then treated with ribonuclease (RNase) H to degrade the original RNA and undergo a subsequent round of PCR to synthesize double-stranded DNA. The double-stranded DNA provides the template for the last *in vitro* transcription (IVT) step, during which biotin-conjugated nucleotides are incorporated into the final complementary RNA (cRNA) product. Samples are then fragmented to reduce secondary structures as well as overall transcript size to

more closely match the 25-base probes. The cRNA samples are then hybridized to a GeneChip and stained with streptavidin-linked phycoerythin. Streptavidin forms a strong bond with the biotinylated nucleotides, attaching the fluorophore phycoerythin to the target transcripts. The sample-hybridized GeneChip is then imaged using an Affymetrix scanner, which quantifies the fluorescence intensity of each pixel and stores the raw data in a DAT file. Affymetrix software is typically used to process the DAT files and calculate probe-level intensity values, which are exported as CEL files. It is with these CEL files that most analyses of Affymetrix GeneChip expression data begin.

2.2.2 Probe summarization

Figure 2.1 provides a visualization of probe-level intensity data read directly from CEL files for three probe-sets targeting high (top), medium (middle) and low (bottom) abundance transcripts, across 33 samples generated for the BXD PFC project. There is a considerable amount of variability observed among probes within a probe-set, despite their targeting the same transcript. This variation is primarily attributed to differences in probe binding affinities, explaining why the relative measurements generated by microarrays are inappropriate for making gene comparisons within arrays (Do et al., 2006).

As is common when working with microarray data, the intensity levels were $\log_2$ transformed prior to plotting. Placing microarray data on a log helps reduces skewness in the distribution, causing it to more closely resemble a normal distribution. This is clearly demonstrated by Figure 2.2, which provides pairwise scatterplots of raw microarray data for three biological replicates both before (top) and after (bottom) log transforming the data. The natural distribution of gene expression levels are not characterized by a normal curve, which would imply most gene are expressed at the

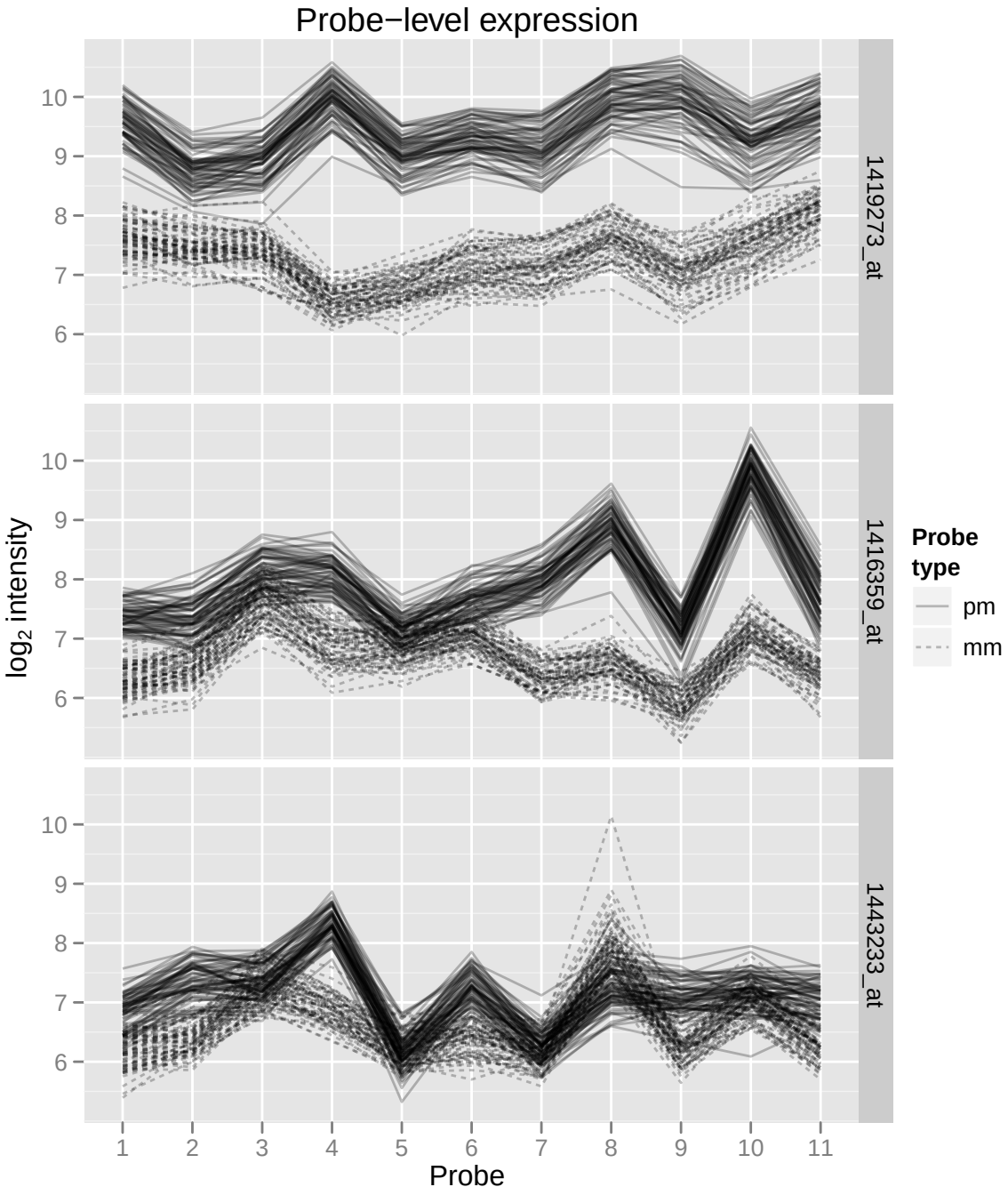

Figure 2.1. Probe-level expression for three probe-sets selected to represent high-, medium- and low-abundance transcripts, respectively. Each line represents a different microarray sample from the BXD PFC dataset. This plot was generated using custom `ggAffy_ProbePlot` function (Source code available in appendix A.2).

same level, with relatively few extreme outliers (Hardin and Wilson, 2009). In fact, a power law may better describe the distribution raw microarray gene expression data (Purdom and Holmes, 2005), which would indicate most genes are expressed at lower levels but a small number of genes expressed at extremely high levels are expected. Still, forcing gene expression data to approximate a normal distribution via log transformation greatly facilitates interpreting differences in transcript abundance by expressing them as fold changes (Quackenbush, 2002).

In order to assess the expression of a target transcript as a whole, the individual probes within a probe-set must be combined. A wide variety of methods for calculating expression summaries exist. The first method to come into widespread use was developed by Affymetrix and provided as part of the Microarray Analysis Suite 4.0 (MAS 4.0). Their summarization strategy involved calculating the Average Difference (*AvgDiff*) in PM/MM intensity for all probe pairs in a set. It is through the subtraction of MM signal that MM probes were intended to fulfill their role and remove background noise. There were issues with this approach, however, namely the *AvgDiff* would occasionally yield negative expression summaries when the intensity of MM probes exceeded their PM counterparts, a biologically inscrutable outcome. Irizarry et al. (2003) observed that approximately 1/3 of MM probe intensities were higher in a dataset comprising five Affymetrix HG-U95 GeneChips. In examining the dataset used to construct Figure 2.2, I found a slightly lower prevalence of MM > PM probe pairs than what Irizarry et al. reported, with 23.7%, 23.5% and 24.2% of probes being affected in the B6_1, B6_2 and B6_3 samples, respectively. Several examples of MM > PM probe-pairs are present in Figure 2.1, especially in the low abundance transcript visualized in the bottom panel.

Affymetrix addressed this issue when they updated from MAS 4.0 to Microarray analysis Suite 5.0 (MAS 5.0). With MAS 5.0, negative expression values were avoided by substituting MM probe data for an 'ideal mismatch' value, which is calculated using

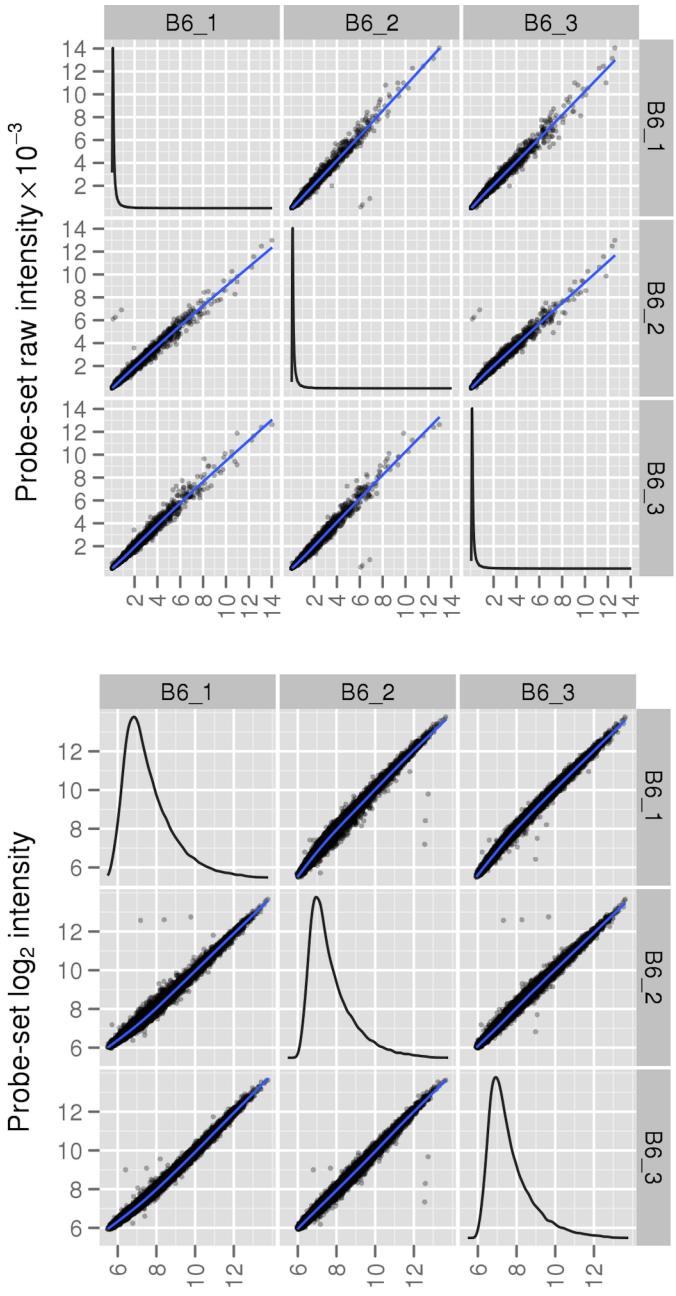

Figure 2.2. Raw intensity versus $\log_2$ transformation Pairwise comparisons of B6 PFC biological replicates using raw probe intensity values (top) and $\log_2$ transformed values (bottom). Density plots along the diagonals depict the distribution of intensities for the sample indicated by the corresponding column label. Intensities are far more evenly spread across the range following $\log_2$ transformation, producing a more normal distribution of gene expression measurments.

PM/MM differences for an entire probe-set, in situations where PM < MM (Affymetrix, 2002). Still, one potential disadvantage of MAS 5.0 is that, like MAS 4.0, expression summaries are calculated for each sample individually, ignoring any information that may be gained through the integration of data across multiple samples. Fortunately, other researchers have developed alternative expression summary approaches that *do* learn with the addition of multiple samples. These model-based approaches leverage the variance across samples, making it possible take into account probe binding affinities (Li and Wong, 2001) and even probe sequence composition (Zhang et al., 2003).

2.2.3 Robust multi-array average (RMA)

The probe expression summarization that is perhaps most commonly used with GeneChip microarrays is the robust multi-array average (RMA), another model-based approach (Irizarry et al., 2003). RMA is especially notable for choosing to ignore MM probes. Irizarry et al. demonstrated that MM probes capture a mixture of non-specific background noise as well as the transcript signal intended exclusively for the PM probes. Irizarry et al. provided a concrete example of how simply ignoring MM probes could benefit down-stream analyses, demonstrating that, when compared to several other expression summaries that rely upon MM subtraction, only the RMA approach could statistically differentiate among probe-sets measuring control transcripts spiked-in at a range of specific concentrations. As such, they argued that any marginal gain in specificity accomplished by subtracting MM probes is not worth the additional noise introduced by this transformation.

Given the strong performance of RMA, it was chosen as this project's primary expression summary for the purpose of measuring absolute transcript abundance. Although the RMA algorithm has been extended to account for probe sequence composition

(Wu et al., 2004), providing marginal improvements in accuracy, particularly for lower abundance transcripts, we chose to stick with the standard RMA approach because it enabled more direct comparisons between our expression data and microarray datasets generated by other laboratories. For example, the majority of BXD expression datasets published on GeneNetwork were summarized using RMA.

2.2.4 Significance score algorithm

A common goal for employing gene expression microarrays is to identify groups of genes that are expressed at different levels across experimental conditions. Achieving this goal typically involves generating microarray data for multiple samples per condition so that traditional statistical tests can be applied. However, the expense and time required to generate microarray data means obtaining a sufficient number of replicate samples is not always practical. This is especially challenging for systems genetics studies, where resources are often used to assay large numbers of different individuals across a population, rather than taking multiple measurements of individual samples.

The S-score algorithm makes it possible to measure differential expression in microarray experiments with few or zero replicate samples (Zhang et al., 2002). Rather than comparing the summarized probe data generated from an expression summary, the S-score approach directly compares probe-level intensities and then combines the relative changes into a single measurement that represents the statistical significance of a gene's change in expression.

The idea of making comparisons across groups of single individuals immediately raises concerns about the reproducibility of any change that is detected. Zhang et al. addressed this concern using microarray expression data for PFC and ventral tegmental area (VTA) from several B6 mice and calculating S-scores within and across brain regions.

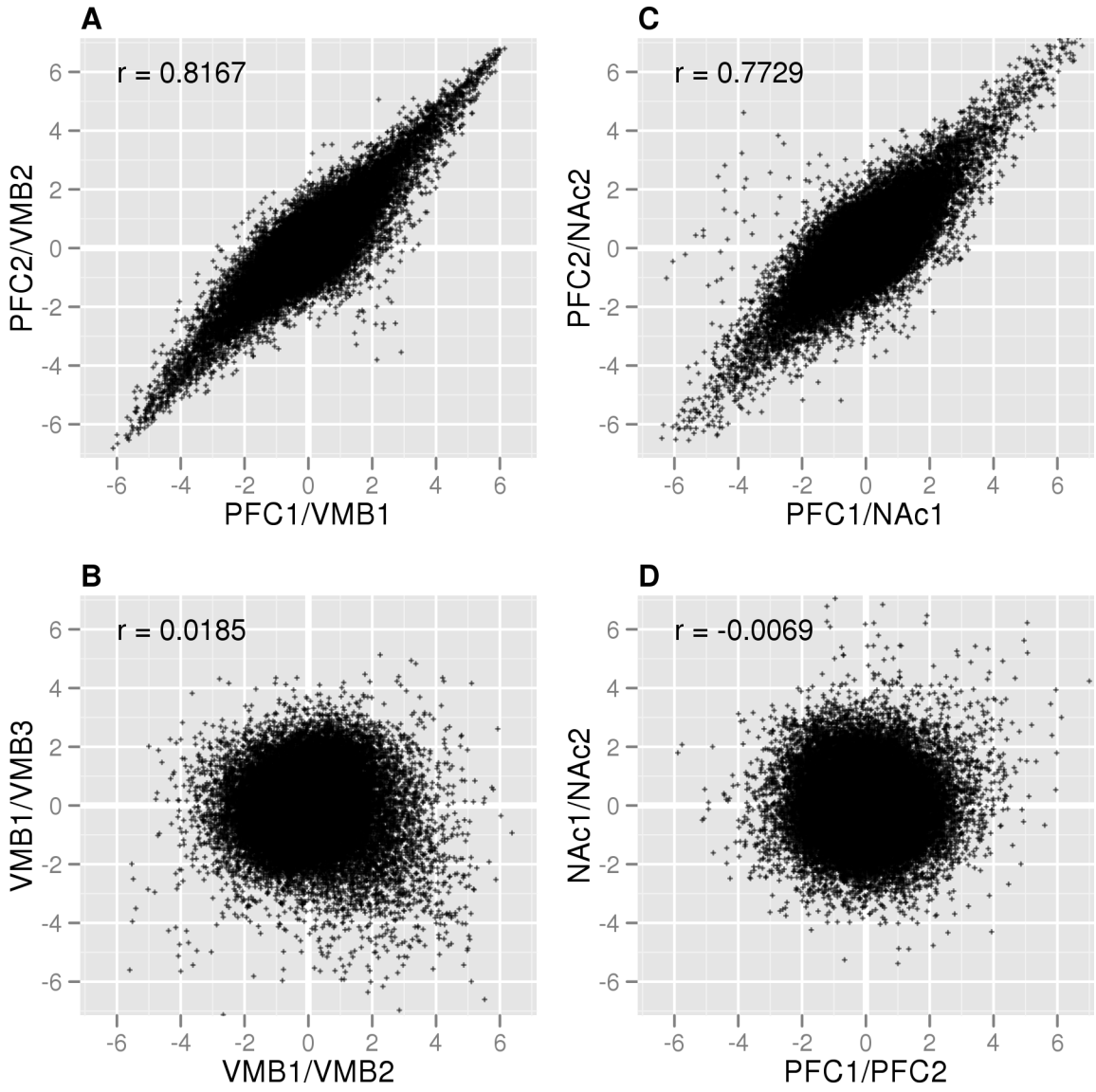

Figure 2.3. S-score reproducibility using M430v2 GeneChips. Two independent comparisons of B6 PFC and VMB expression profiles detected by S-scores were highly reproducible (A), whereas S-scores generated for pairwise comparisons of B6 VMB showed no correspondence and were randomly distributed around zero (B). A similar pattern of results was achieved by comparing independent S-score analyses of PFC/NAc differences (C) and within-region differences (D). Collectively, these analyses replicated the results published by Zhang et al. (2002) and Kerns et al. (2003).

If the differentially expressed genes detected by S-scores reflect genuine differences in brain region transcriptomes they should be reproducible across biological replicates, whereas differences detected within-regions should be attributable to random error and inconsistent across biological replicates. Indeed, Zhang et al. (2002) found a significant correlation between replicate PFC/VTA comparison experiments ($r = 0.75$), while the correlation between replicate VTA/VTA comparisons was effectively null ($r = 0.17$).

This experiment had been carried out using Affymetrix Mu6500 GeneChips: one of the earliest Affymetrix GeneChip models. To ensure the outcome of this analysis could be replicated with the updated M430v2 GeneChips used here, which measure more than 6× the number of transcripts targeted by the Mu6500 GeneChip, I repeated the analysis outlined by Zhang et al. (2002). However, because we collected only 3 replicates for each progenitor strain, it wasn't possible to conduct two independent within-region differential expression analyses. Still, the correlation between the VMB/VMB S-scores was effectively zero ($r = 0.019$), despite the fact that sample VMB1 was used in both comparisons (Figure 2.3 B). Importantly, the correlation between replicate PFC/VMB comparisons was highly significant ($r = 0.82$) across all M430v2 probe-sets (Figure 2.3 A). If all probe-sets with $|\text{S-scores}| \leq 2$ are thrown out, that is, probe-sets whose differences fall within the margin of random error, the correlation jumps to 0.98, clearly demonstrating that reproducible changes can be detected using S-scores to compare single samples.

Kerns et al. (2003) performed a similar validation study, except PFC expression was compared with NAc, rather than VMB, and they used the updated Affymetrix Mouse Genome U74 2.0 (U74v2) GeneChips, which measure twice as many transcripts as Mu6500 microarrays, for a total of 12,488 probe-sets. Similar to what was reported by Zhang et al. (2002), they found that cross-region S-scores were highly reproducible between repeat experiments ($r = 0.81$), while the correlation between S-scores calculated

within PFC and NAc correlated poorly ($r = 0.004$). Here too, I repeated this analysis using M430v2 GeneChip data and was able to reproduce their results (Figure 2.3 C–D). Kerns et al. (2003) also identified genes differentially expressed between PFC and NAc using MAS 5.0 so the results could be compared with those obtained using S-scores. The two methods largely agreed about which genes were significantly different between brain regions, with the notable exception of an outlier group comprised of probe-sets with large MAS 5.0 fold-changes but S-scores ≈ 0. Many of these outlier probe-sets had been deemed 'absent' by the MAS 5.0 detection-calls algorithm and thus represented low quality measurements that were effectively filtered out through the S-score analysis (Kerns et al., 2003).

Molecular validation of gene expression differences detected using S-scores was provided by a microarray study seeking to identify common molecular mechanisms underlying normal neurogenesis associated with dentate gyrus (DG) development and aberrant neurogenesis that characterizes temporal lobe epilepsy (Elliott et al., 2003). After identifying 37 genes differentially expressed under both neurogenic conditions, Elliott et al. performed *in situ* hybridization using labelled riboprobes to measure transcript levels for 17 of these genes across frozen coronal brain sections. The *in situ* evidence for 12 of these genes confirmed the results from the original S-score analysis. Further molecular validation was provided in a study published by Kennedy et al. (2006a), which utilized the same spike-in dataset used by Irizarry et al. (2003) to validate the RMA approach. Again, the goal is to successfully distinguish between control probe-sets, which have been spiked-in at known concentrations, and all other probe-sets. They found that both S-scores and RMA values were able to isolate the spiked-in probe-sets, although S-scores could do so at lower concentrations than RMA (Kennedy et al., 2006a). MAS 5.0, on the other hand, had difficulty distinguishing between control and background probe-sets at any concentration.

While the S-score algorithm has proved capable of identifying biologically meaningful changes in gene expression, Zhang et al. noted that a significant change identified by a large S-score does not necessarily imply a reproducible change. Although S-scores are calculated using multiple probe-pair comparisons, probes target different transcript regions[1] and thus do not constitute replicate measurements. It is only through the inclusion of biological replicates that a significant change in gene expression can be deemed reproducible and not artifactual. Still, the studies in this section have clearly demonstrated that S-scores provide a highly sensitive measure of differential expression, capable of detecting both reproducible and biologically meaningful changes, assuming the microarray expression data under consideration is clean and of high quality. As such, the S-score algorithm was used as this project's primary mean of measuring differential gene expression between saline and ethanol treated samples.

2.3 Microarray data generation

2.3.1 Animals and tissue collection

B6 and D2 strains and BXD RI strains 1–42 were purchased from Jackson Laboratory (Bar Harbor, ME). The novel BXD strains were derived from the independent advanced intercross (AI) were acquired from Oak Ridge National Laboratory (Oak Ridge, TN, USA). All animals were male and between 10–12 weeks of age. Mice were housed 4 per cage with *ad libitum* access to standard rodent chow (catalog #7912, Harlan Teklad, Madison, WI) and water. Following a two week acclimation period mice were injected intraperitoneal (IP) with saline or 1.8 g/kg of ethanol. This ethanol dose was originally chosen from pilot experiment data to maximize anxiolytic activity and minimize sedative responses (decreased locomotor activity) as part of the ethanol-induced anxiolysis study

presented in Chapter 4. For those experiments, all mice underwent behavioral testing that included 15 minutes of restraint in a 50 mL conical tube followed by 10 minutes in a light-dark chamber. The results of these behavioral genetics experiments are discussed Chapter 4. Mice were killed by cervical dislocation four hours following IP injection. Immediately thereafter, brains were extracted and chilled for one minute in iced phosphate buffer before being microdissected into 8 constituent regions as described previously (Kerns et al., 2005), including PFC, NAc and VMB, which includes VTA and substantia nigra. Excised regions were placed in individual tubes, flash-frozen in liquid nitrogen and stored at $-80°$C.

Experimental procedures were approved by Virginia Commonwealth University Institutional Animal Care and Use Committees in accordance with the National Institutes of Health.

2.3.2 Microarray data generation

This study incorporated PFC tissue from 27 BXD strains, NAc and VMB tissue from 35 BXD strains, as well as B6 and D2 tissue from all three regions. The complete sample inventory is provided in Table 2.1. Frozen tissue for a given brain region and strain was pooled from 4–5 animals and homogenized with Aurum^{TM} total RNA fatty and fibrous tissue extraction kit (BioRad, catalog #732-6830) and a Tekmar homogenizer. RNA concentration was determined by absorbance at 260 nm, and RNA quality was analyzed by electrophoresis with an Experion analyzer (BioRad, Hercules, CA) and 260/280 absorbance ratios. All RNA samples had RNA quality indices (RQI) ≥ 8. Total RNA ($5\mu g$) derived from each pool and spike-in poly-A RNA controls were reverse transcribed into double-stranded cDNA using Affymetrix SuperScript® one-cycle cDNA kit (Invitrogen, catalog #A10752030). Biotin-labeled cRNA was synthesized from cDNA using the

Table 2.1. Microarray sample inventory

	Prefrontal cortex		Nucleus accumbens		Ventral midbrain	
	Saline	Ethanol	Saline	Ethanol	Saline	Ethanol
B6	3	3	3	3	3	3
D2	3	3	3	3	3	3
BXD01	1	1	1	1	1	1
BXD02	1	1	1	1	1	1
BXD05	1	1	1	1	1	1
BXD06	1	1	1	1	1	1
BXD08	1	1	1	1	1	1
BXD09	1	1	1	1	1	1
BXD11	1	1	1	1	1	1
BXD12	1	1	1	1	1	1
BXD14	1	1	1	1	1	1
BXD15	1	1	1	1	1	1
BXD16	1	1	1	1	1	1
BXD18	1	1	1	1	1	1
BXD19	1	1	1	1	1	1
BXD20	1	1	1	1	1	1
BXD21	1	1	1	1	1	1
BXD22	1	1	1	1	1	1
BXD23	1	1	1	1	1	1
BXD27	1	1	1	1	1	1
BXD28	1	1	1	1	1	1
BXD31	1	1	1	1	1	1
BXD32	1	1	1	1	1	1
BXD33	1	1	1	1	1	1
BXD34	1	1	1	1	1	1
BXD36	1	1	1	1	1	1
BXD38	1	1	1	1	1	1
BXD39	1	1	1	1	1	1
BXD40	0	0	1	1	1	1
BXD42	1	1	1	1	1	1
BXD43	0	0	1	1	1	1
BXD48	0	0	1	1	1	1
BXD63	0	0	1	1	1	1
BXD66	0	0	1	1	1	1
BXD67	0	0	1	1	1	1
BXD90	0	0	1	1	1	1
BXD98	0	0	1	1	1	1

Number of microarrays processed for each brain region and treatment group.

GeneChip IVT labeling kit (Affymetrix, part #900449) according to manufacturer's instructions, purified using the RNAeasy Mini Kit (Qiagen, Mountain View, CA), and quantified by absorbance at 260 nm. Labeled cRNA samples were hybridized to M430v2 microarrays (Affymetrix, part #900497) according to the manufacturer's protocol.

2.3.3 Mitigating batch effects

The number of microarrays involved in this study required that their processing be divided into smaller groups of manageable sizes. For large-scale experiments such as this that require samples to be processed in subsets, special attention must be paid to prevent introducing batch effects: that is, non-biological gene expression variation produced by the systematic grouping of samples throughout the protocol. Generating microarray expression data involves many steps, each of which has the potential to introduce non-biological expression heterogeneity. Supervised randomization techniques at every stage in the protocol can help ensure that changes in gene expression are produced by the experimental variables of interest, rather than a byproduct of technical factors. If unaccounted for, batch effects can majorly impact the results of an experiment. For example, Lamb et al. performed a large-scale microarray experiment that sought to systematically characterize the effects of small molecule drugs on the expression of different human cell lines and reported that hierarchical clustering primarily grouped samples by cell type and cell culture batch, obscuring the effects of drug treatment (Lamb et al., 2006).

While a number of methods exist that attempt to correct batch effects (Alter et al., 2000; Benito et al., 2004; Johnson et al., 2007), it is always preferable to avoid the need for such corrections in the first place using experimental design strategies that mitigate batch effects. To that end, we performed a supervised randomization of samples into

different batch groups prior to each of the following microarray processing stages: total RNA extraction, cRNA synthesis and sample hybridization. However, to minimize the risk of technical variation confounding expression variation driven by ethanol, both a saline and ethanol-treated mouse from a single strain were always processed together.

2.3.4 Microarray quality assessment

Microarray data quality was assessed by inspecting the distributions of log-transformed probe intensity values, as well as scanning for outlier chips using a standard battery of quality measurements, including: average background, scaling factor, percentage of probe-sets called present and $3'/5'$ ratios for *Actb* and *Gapdh*. Relevant quality assessment figures are provided in the appendix.

Bioconductor's implementation of the MAS 5.0 Detection Calls Algorithm, available in the `affy` package (Gentleman et al., 2004) for R, was used to generate absent present marginal calls across all samples. We excluded any probe-sets called absent in $\geq$ 95% of samples from all subsequent analyses to improve the ratio of true positives in downstream statistical filtering (McClintick and Edenberg, 2006). This removed 14,096, 12,970 and 13,312 probe-sets from the PFC, NAc and VMB, respectively. The lists of 'absent' probe-sets were largely overlapping, with 11,343 probe-sets filtered out of all 3 regional datasets, suggesting this filtering step largely removes probe-sets targeting genes unexpressed in brain tissue.

Expression data from the saline and ethanol treatment groups were background corrected, quantile normalized and summarized using the RMA expression measure (Irizarry et al., 2003). All datasets generated for this paper can be queried on GeneNetwork or downloaded in their entirety in a minimum information about a microarray experiment (MIAME) compliant form from the Gene Expression Omnibus repository

under accession number GSE28515. Microarray probe-sets can be annotated from a variety of sources. Annotation data for Affymetrix M430v2 probe-sets was obtained from the GeneNetwork Data Sharing Zone. Here GeneNetwork was chosen because it was consistently a more complete dataset that was updated with greater frequency.

2.4 Differential expression analysis

The large scale of this study made cost prohibitive the inclusion of biological replicates for each RI strain across treatment groups. Therefore, assessing the reproducibility of changes in gene expression within a single strain by conventional methods, such as the Significance Analysis of Microarrays (SAM) approach (Tusher et al., 2001), was not possible. We used an alternative approach to identify probe-sets with extreme ethanol expression changes across a minority of strains or smaller but consistent changes across a larger portion of the BXD family. The impact of acute ethanol on transcript abundance was measured using the S-score algorithm (Zhang et al., 2002), which utilizes probe-level data to determine the statistical significance of transcript level differences between a pair of Affymetrix microarrays. We utilized the R implementation of the S-score algorithm (Kennedy et al., 2006b) to compare microarray expression levels within BXD strains across treatment groups to generate a saline versus ethanol S-score for each probe-set, where a positive S-score indicates up-regulation with ethanol and vice-versa. In the case of the B6 and D2 progenitor strains, where biological replicate microarrays were available for each strain in triplicate, S-scores were generated using the `SScore` function's `classlabel` argument.

Statistical significance of a given probe-set's ethanol response across strains was assessed using the approach proposed by Fisher (1925) for combining p-values from

multiple tests:

$$S = -2\sum_{i=1}^{n}\log(p_i) \tag{2.1}$$

where n is the number of tests and p corresponds to the ith p-value. Each probe-set S-score for a single strain was considered an independent test. S-scores are normally distributed with a mean of 0 and a standard deviation (SD) of 1 (Zhang et al., 2002). For 2-tailed tests, p-values for each probe-set were calculated as twice the probability of obtaining an S-score at least as large as the absolute value of the observed S-score. This calculation was rendered in R using the following code:

```
2 * pnorm(abs(x), lower.tail = FALSE)
```

where x is the original S-score. Equation 2.1 was used used to combine the S-score transformed p-values. This process was then repeated for 1,000 random permutations of the observed S-score expression matrix, so that empirical p-values could be obtained by comparing observed results to the permutation distribution. Finally, to correct for multiple testing, q-values were generated from the empirical p-values (Benjamini and Hochberg, 1995). Probe-sets with q-values ≤ 0.05 were considered to be significantly ethanol responsive. This analysis has been implemented as a R function, the source code for which is provide in appendix A.1.

2.4.1 Ethanol responsive genes across BXD panel

Kerns et al. previously reported an initial microarray analysis of PFC, NAc and VMB brain regions from the B6 and D2 inbred strains and identified 307 genes that changed significantly with acute ethanol treatment (Kerns et al., 2005). To extend those prior efforts and observe the genetic correlations that exist among ethanol sensitive genes, we performed a similar analysis using PFC, NAc, and VMB expression data obtained

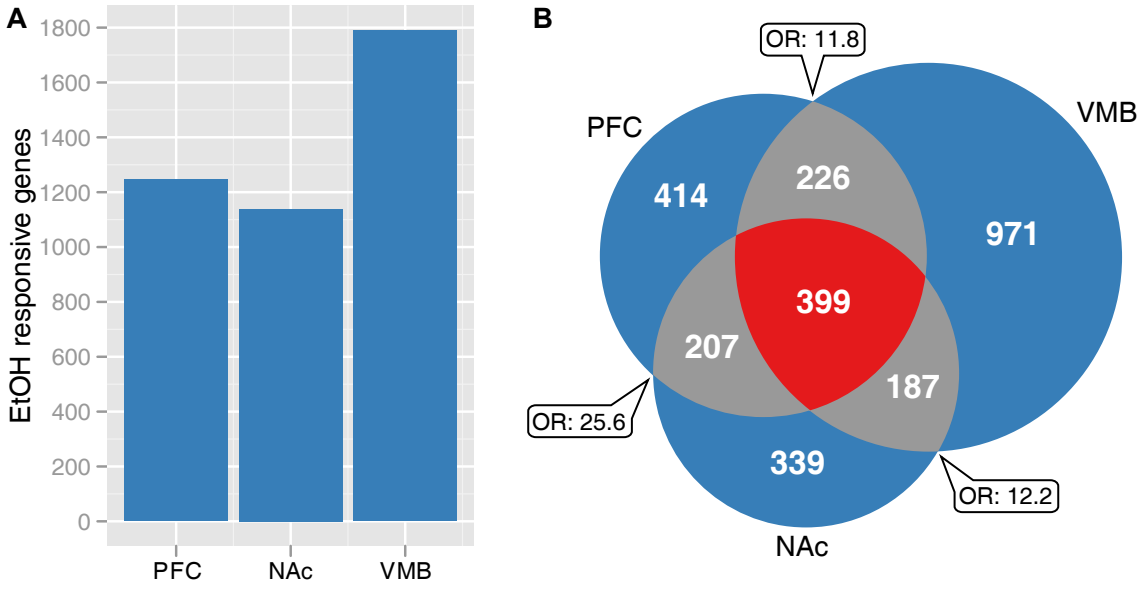

Figure 2.4. Transcriptional response to acute-ethanol within 3 regions of meso-corticolimbic reward circuit across the BXD family A. Number of genes found to be significantly ethanol-responsive in the PFC (n=29), NAc (n=37) and VMB (n=37) by analysis of saline vs ethanol S-scores across BXD, B6 and D2 samples. **B.** Venn-diagram depicting which subsets of ethanol-responsive genes are region specific (blue), overlap across two regions (grey) or common to all three regions (red). All three pairwise overlap combinations were statistically significant as determined by Fisher's Exact Test for count data. Odds ratios from this analysis are depicted in word bubbles.

from the BXD family, as well as the founding B6 and D2 strains. The greater genetic diversity provided by the BXD microarray data made it possible to detect gene expression differences that would otherwise be absent in a study limited to the B6 and D2 strains due to epistatic suppression.

As described in the Differential expression analysis section, we used the S-score algorithm for probe-level analysis of each strain's transcriptional response to ethanol, followed by Fisher's combined probability test. This approach favors genes that consistently responded to ethanol across numerous BXD strains, regardless of direction, rather than genes that exhibited large differences in only a small subset of strains. Analysis of microarray datasets for PFC, NAc and VMB identified 3,512 probe-sets, corresponding to 2,743 unique genes, that changed significantly with ethanol in at least one brain region (Table S1). These results replicated over 40% of the genes previously identified as ethanol-responsive by Kerns et al. (2005), despite differences in microarray design, investigators and analysis methods. VMB exhibited the largest transcriptional response to ethanol, while changes observed in PFC and NAc were of comparable magnitude (Figure 2.4A). The transcriptional response to ethanol within each brain region included both unique and shared gene components. Roughly 1/3rd of significantly ethanol-responsive genes in the PFC and NAc were unique to their respective regions, while greater than 50% of the VMB ethanol profile was specific to that region (Figure 2.4B).

2.4.2 Ethanol responsive transcriptional profiles

Assaying gene expression across the BXD panel allowed us to analyze how genetic variation influenced transcriptional responses to ethanol (Figure 2.5). As seen with other heritable complex traits measured in genetic mapping panels, the transcript-level response of most ethanol sensitive genes followed a continuous distribution across the

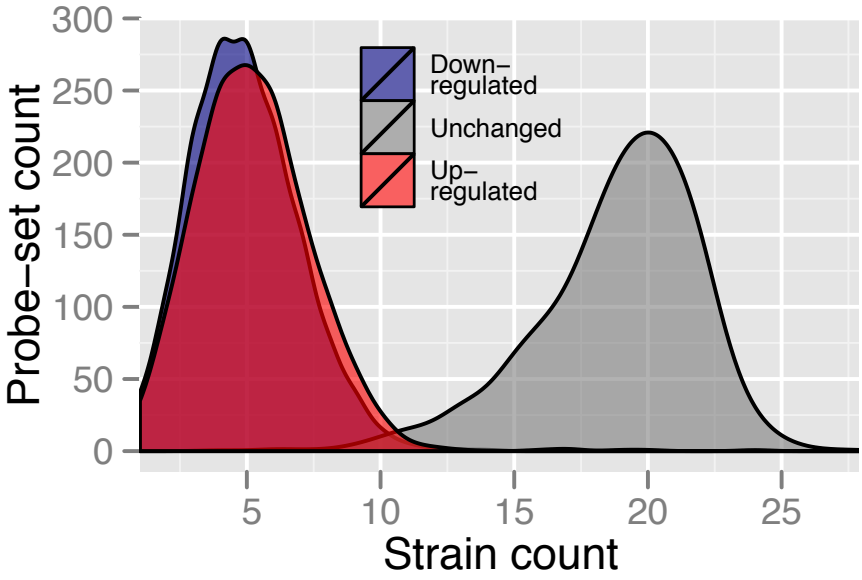

Figure 2.5. Frequency of ethanol responsive classes. Strain frequency distributions of gene transcriptional-response classes based on PFC S-score analysis. S-scores > 2 indicate a gene was up-regulated by acute ethanol, S-scores < 2 indicate down-regulation and S-scores between these thresholds were considered unchanged.

BXD and progenitor strains.

There was a subset of genes that were almost uniformly up-regulated by ethanol, including *Npas4*, *Fos*, *Hsp8*, *Egr2*, *Dusp1* and *Jun*, all of which are neuronal activity dependent. Most genes, however, exhibited divergent ethanol responses between variable subsets of BXD strains, as can be clearly seen by viewing the S-score distributions of the mostly ethanol responsive genes in the PFC (Figure 2.6), NAc (Figure 2.7) and VMB (Figure 2.8).

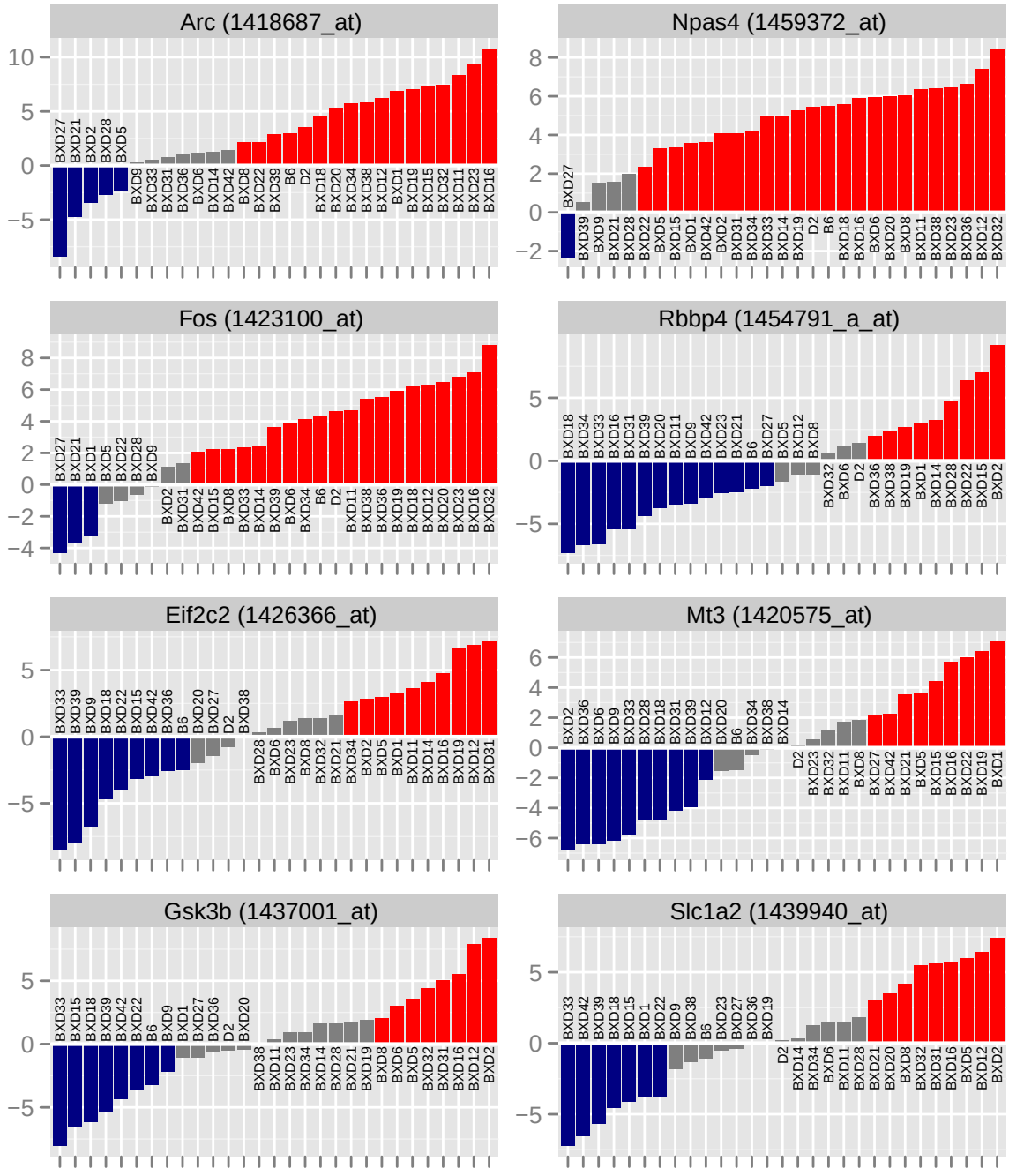

Figure 2.6. Top ethanol responsive genes in PFC S-score distributions for the 8 genes with the most robust transcriptional response to acute ethanol in the PFC.

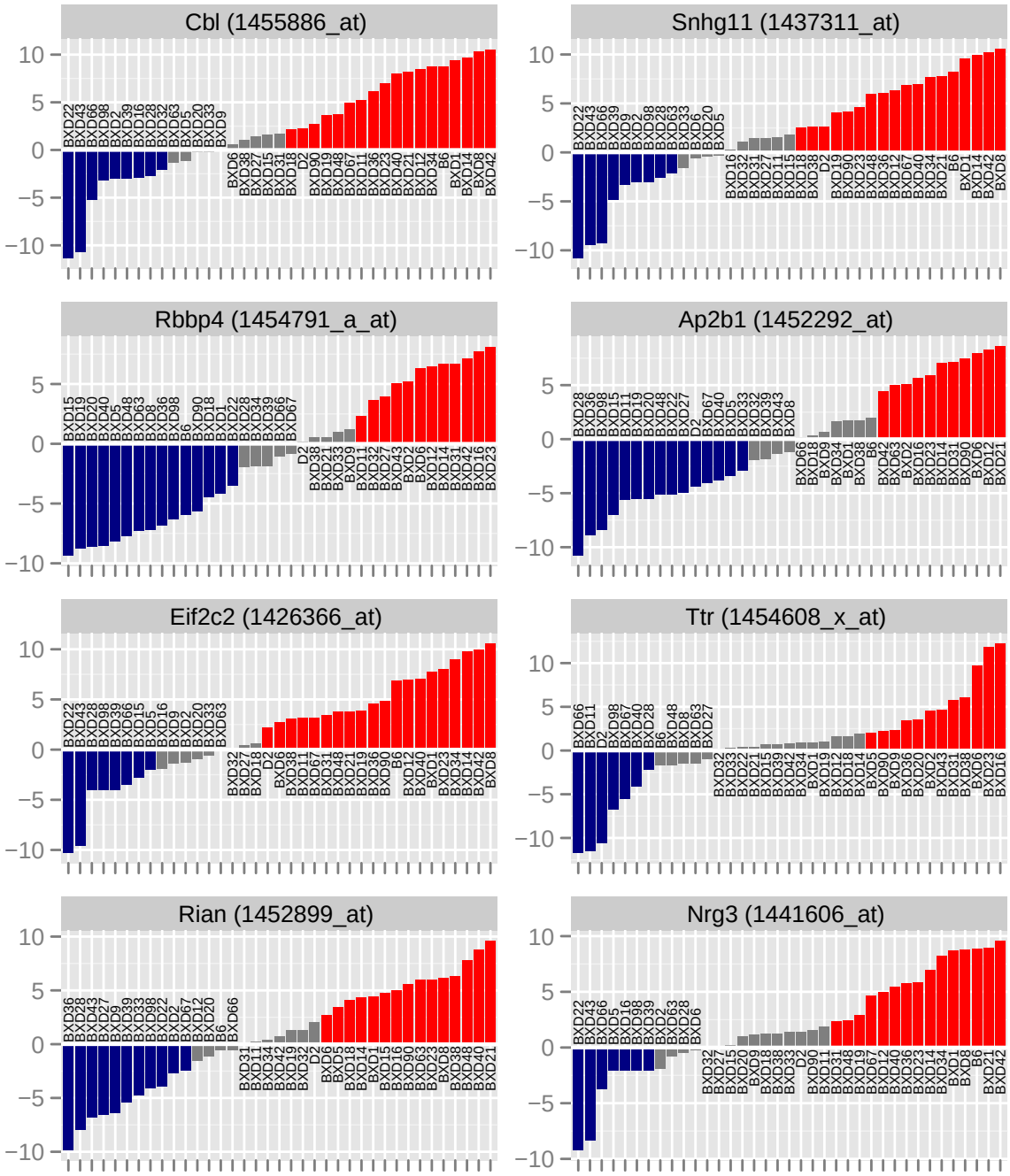

Figure 2.7. Top ethanol responsive genes in NAc

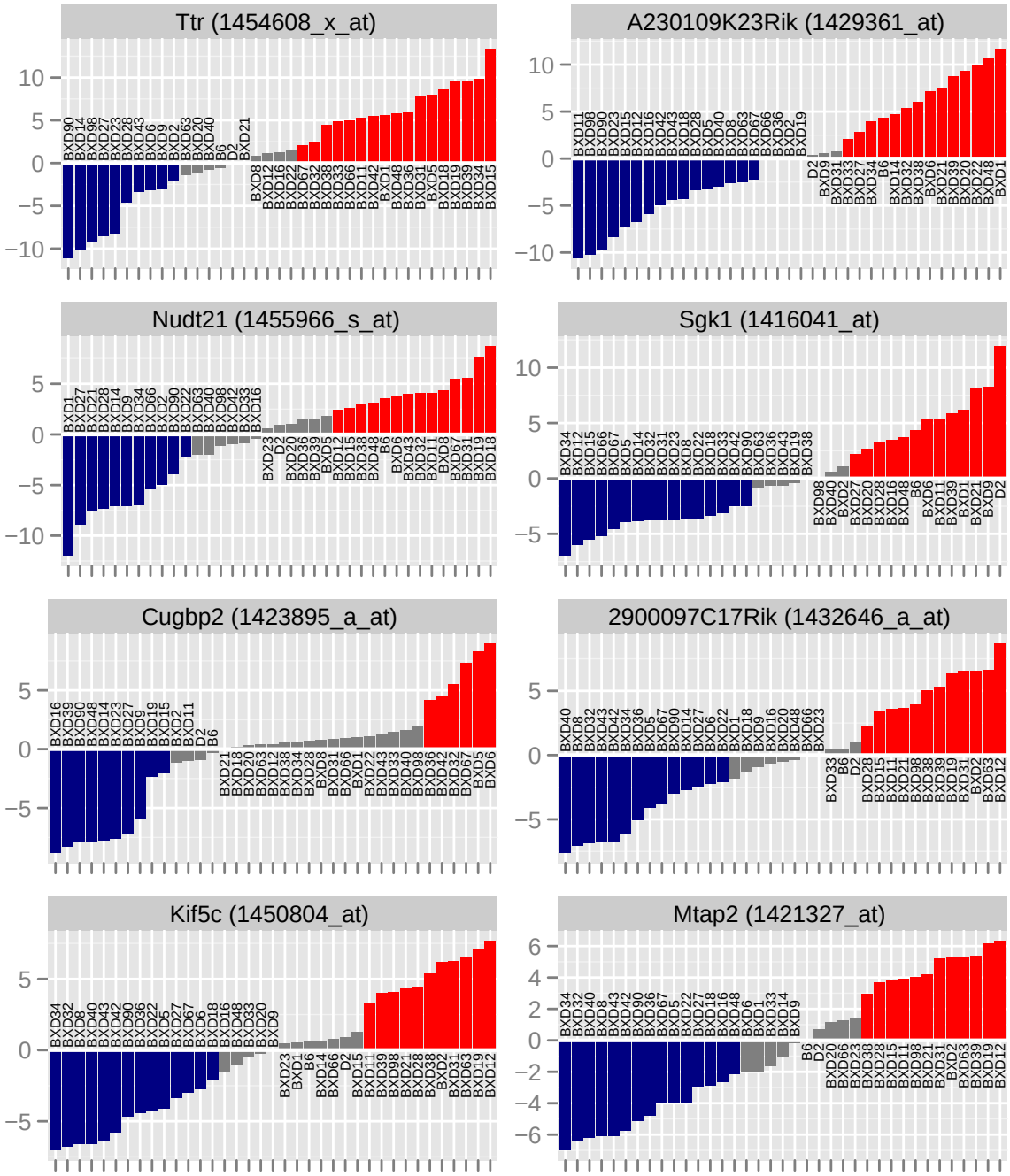

Figure 2.8. Top ethanol responsive genes in VMB

2.4.3 Ethanol responses across brain regions

While continuous distributions of transcriptional responses to ethanol were observed in all profiled regions, the transcript-level changes were highly region specific. Calculating within-gene correlations using S-scores revealed little correspondence between a gene's ethanol response across regions. Even when the scope of this analysis was limited to the 399 genes found to be significantly ethanol responsive in all three brain regions, cross-region S-score correlations were effectively null for all but a small subset of genes. Therefore, acute ethanol effects on gene expression are likely mediated through an interaction between genetic background and brain region specific environmental factors.

However, with sample sizes of 27 in the PFC comparisons and 33 in the NAc/VMB comparison, our statistical power was only sufficient to detect correlations greater than 0.52 and 0.47, respectively, at a significance level of 0.05. Therefore, we were unable to detect the presence of weaker correlations that may exist between these regions. Furthermore, as this analysis was conducted using correlation coefficients, only linear relationships between genes could be discovered. Therefore, using this approach we were unable to address whether nonlinear interactions existed across brain regions in our datasets. Given that nonlinear gene expression relationships have been observed (Kholodenko et al., 1999), and methods exist for constructing gene networks that incorporate nonlinear interactions (Zhu et al., 2007), this may be a possibility that should be addressed in the future.

The small contingent of genes that did exhibit coordinated ethanol responses across regions are listed in Table 2.2. Among these probe-sets, significant cross-regional correlations existed between the PFC and NAc; not between PFC & VMB or NAc & VMB. According to the latest Affymetrix M430v2 annotations, only 3 of the 5 probe-sets target protein-coding genes; `1427747_a_at`targets *Lcn2*, `1418687_at`targets *Arc* and

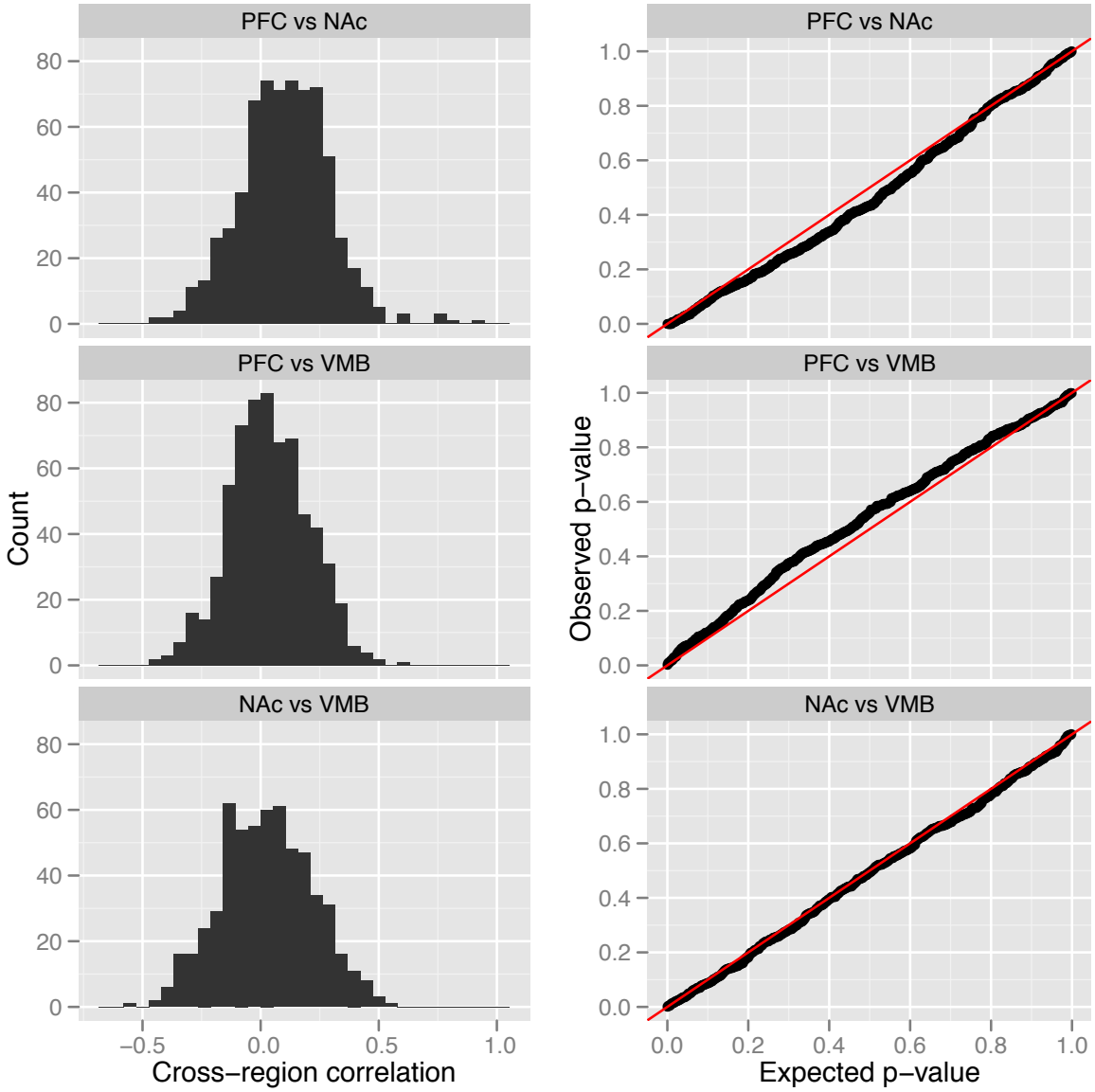

Figure 2.9. Cross-region correlations of ethanol responsive gene expression. In order to determine the degree to which a gene's transcriptional response to ethanol is tissue-specific, we calculated cross-regional S-score correlations for each of the 399 probe-sets that were differentially expressed in the PFC, NAc and VMB. Histograms on the left display the distribution of Pearson correlation coefficients for these 399 genes. Plotting p-values from this analysis against a uniform distribution indicate there is effectively no coordinated response to ethanol across regions.

1416505_attargets *Nr4a1*. Scatterplots in Figure 2.10 compare PFC S-scores for these 3 probe-sets against their S-scores in the NAc.

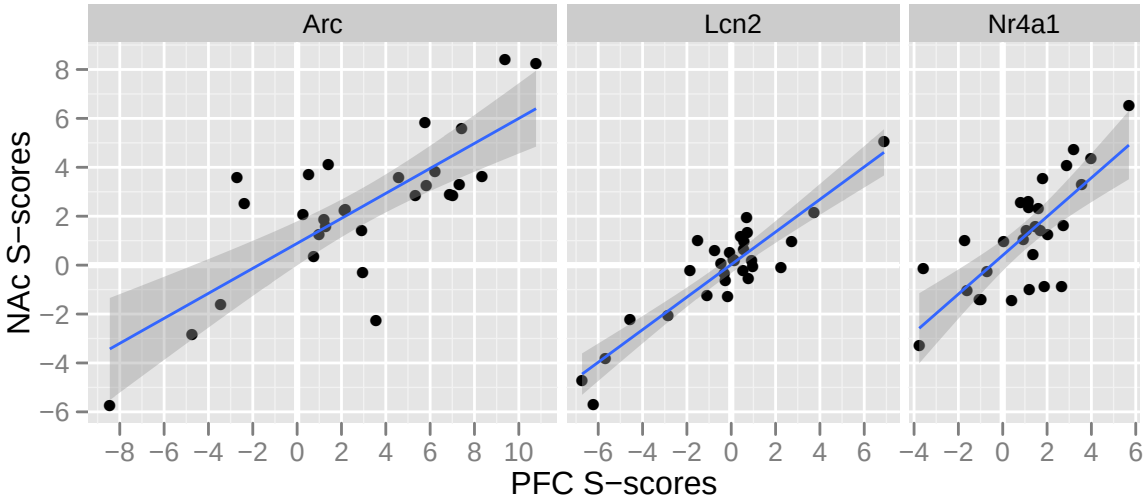

Figure 2.10. Coordinated ethanol responses across PFC and NAc Scatterplots for the probe-sets from Table 2.2 that target known protein-coding genes. The blue lines represent fitted linear models, with 95% CIs represented by the grey bands.

Table 2.2. Coordinated ethanol responses across PFC and NAc.

Probe-set	Gene	*r*	p-value
1427747_a_at	*Lcn2*	0.91	1.14e-11
1427820_at	BC021831	0.79	2.94e-07
1418687_at	*Arc*	0.76	1.60e-06
1416505_at	*Nr4a1*	0.75	2.53e-06
1440342_at	BB276544	0.74	3.79e-06

List of probe-sets whose PFC S-scores were significantly correlated with their NAc S-scores, calculated using Pearson's correlation coefficient (r). No significant correlations were observed between PFC & VMB or NAc & VMB. Scatterplots for several of these probe-sets are provided in Figure 2.10.

2.4.4 Functional analysis of ethanol responsive genes

Functional enrichment analyses were performed using ToppFun, a functional enrichment application available at toppgene.cchmc.org as part of the ToppGene suite of web applications (Chen et al., 2009b). Entrez ID's were submitted and analyzed for over-representation of genes that belong to a GO category (cellular component, molecular function and biological process), biological pathway, gene family or, similarly, encode a particular protein domain. In order to enhance the specificity and informativeness of these results, we considered only those categories that comprise greater than 3 and fewer than 300 genes, inclusive. Multiple testing was accounted for using a 1% false discovery rate (FDR) threshold. Results were curated by excluding categories with gene lists more than 80% redundant with other, less enriched, categories.

Functional enrichment analysis showed strong homology in the functional categories regulated by ethanol in all three regions (Table S2). Gene groups related to synaptic activity and plasticity were among the most significantly over-represented GO biological functions, with dendritic or synaptic structure as the top GO cellular component (CC) in each region. The 399 genes that were significantly ethanol-responsive in all three brain regions were also highly enriched for proteins that localize to the pre- and post-synaptic membranes and regulate synaptic transmission, including both ionotropic and metabotropic glutamate receptor categories. However, there were notable regional differences; for example, the over-representation of GABA and glutamate receptor signaling pathways was particularly high in VMB.

2.5 Overlap with other AUD-relevant microarray data

The approach used here to identify ethanol-responsive genes was somewhat unorthodox for a microarray study. Rather than comparing two treatment groups composed of multiple biological replicates, our treatment groups comprised relatively large samples of 29–37 genetically unique strains. Although only single arrays were performed per strain/treatment, the issue of biological variability was mitigated by pooling tissue samples from at least 3 biological replicates per strain. Despite these methodological differences our results largely agreed with more traditional ethanol-relevant microarray analyses of brain gene expression. For example, our analysis replicated a large contingent of the genes found to be differentially expressed by acute ethanol in Kerns et al.'s original study of PFC, NAc and VTA expression across B6 and D2 mice, which did include multiple replicates of each inbred strain (2005).

Concurrently with the BXD project, the Miles laboratory generated a similar microarray expression dataset for the LXS RI panel. This dataset included PFC expression data for two sets of 43 LXS strains, each receiving an IP injection of either saline or ethanol (2 g/kg). I repeated with the LXS data the analyses outlined above for the BXD expression data, including generating saline versus ethanol S-scores and repeated the differential expression analysis described in section 2.4. This identified 1,811 probe-sets, targeting 1,482 unique genes, that were significantly responsive to acute ethanol across the profiled LXS strains. The full list of genes is provided in Table S3. The lists of significantly ethanol responsive genes identified in the PFC of BXD and LXS overlapped extensively (Figure 2.11).

We also observed significant overlap with genomic studies that have investigated other aspects of AUDs. At the other end of the AUD spectrum are studies concerned with the impact of chronic ethanol consumption, particularly as it relates to stress-induced

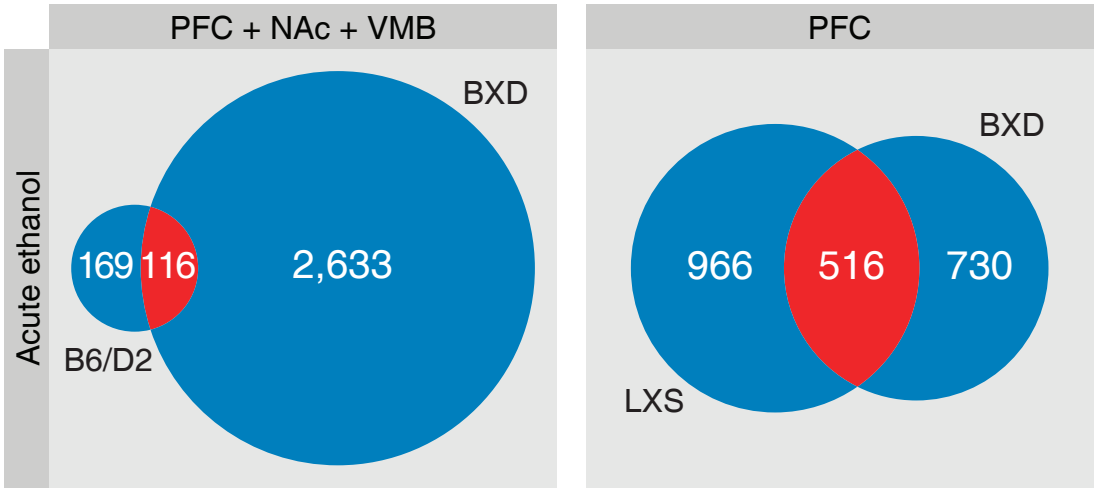

Figure 2.11. Overlap with other models of acute ethanol.

consumption. As the stress and anxiety associated with alcohol withdrawal are thought to be key drivers of ethanol drinking in alcoholics (Koob, 2003). A well characterized mouse model of this paradigm, referred to as chronic intermittent ethanol exposure (CIE), uses repeated cycles of ethanol exposure via vapor chambers to induce dependence. Over time, mice that undergo this procedure exhibit an enhanced preference for ethanol in two-bottle choice experiments (Becker and Lopez, 2004). In collaboration between Becker and Lopez and the Miles laboratory, a set of BXD strains were subjected to this protocol and used to generate a novel CIE microarray data resource. As with our acute ethanol microarray expression data, this resource was made up of two treatment groups; in this context, mice either belonged to the ethanol vapor treatment group or the air control group. I used the data from these two groups to generate S-scores and repeated the same differential expression analysis. Here too, we observed substantial overlap with our acute ethanol BXD results Figure 2.12.

The study by Mulligan et al. (2006), described in section 1.3.2, used microarray data from whole brain RNA to perform a meta-analysis across several inbred lines used as models of high and low ethanol consumption. Nearly a quarter of the BXD ethanol-

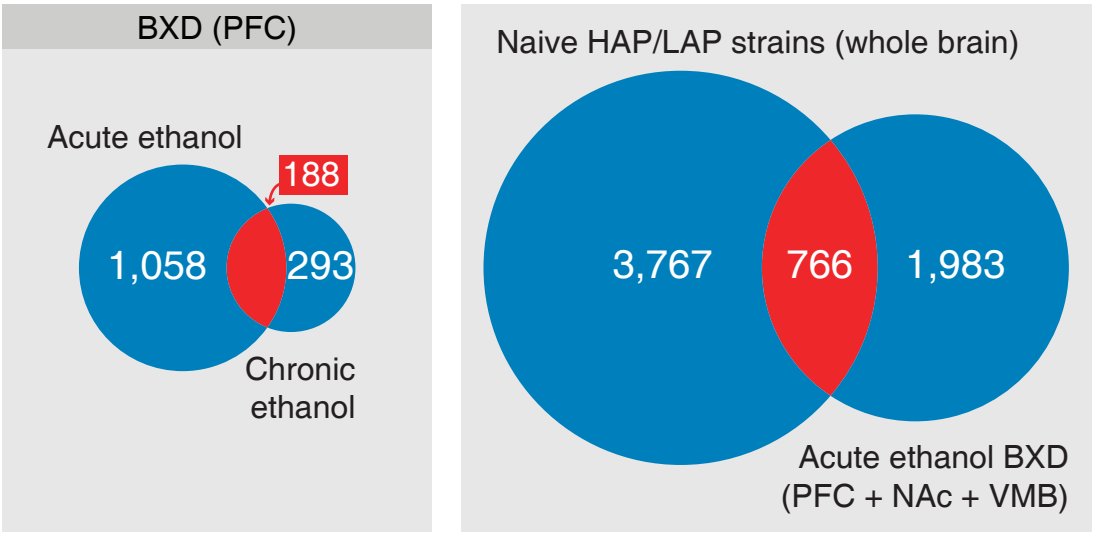

Figure 2.12. Overlap with models of other aspects of AUD.

responsive genes defined here were among the list of genes with basal expression levels found to significantly co-vary with ethanol preference (Figure 2.12). The extent of this overlap might have been greater if the meta-analysis had been conducted across targeted brain regions, rather than whole brain. Regardless, many of the genes whose basal expression levels segregate with alleles driving divergent preferences for ethanol were also regulated upon exposure to acute ethanol in our study.

2.6 Discussion

Using an extensive microarray gene expression dataset, comprising PFC, NAc and VMB transcriptional profiles for a subset of the BXD RI panel (Table 2.1), we have investigated the genetic consequences of acute ethanol exposure in the mesocorticolimbic pathway. The BXD mice were derived from the B6 and D2 inbred strains, which exhibit divergent responses in a number of ethanol-relevant phenotypic assays (Belknap et al., 1993; Grieve and Littleton, 1979; Metten and Crabbe, 1994; Phillips et al., 1995). Similar to

what was observed in the microarray study of B6 and D2 mice performed by Kerns et al. (2005), we identified a robust transcriptional response acute ethanol across all three assayed brain regions.

While many exhibited region-specific responses, a subset of 399 genes were jointly regulated by acute ethanol in all three brain regions (Figure 2.4). Functional analysis of this group indicated an over-representation of genes that regulate synaptic transmission (p-value = 1.8E-09 (Table S2). This included three potassium voltage-gated channels and *Kcnma1*, the α1 subunit of the Ca^{+}-activated BK K^{+} channel. *Kcnma1* directly regulates acute ethanol sensitivity in *C. elegans* (Davies et al., 2003) and is a major hub within one of the ethanol responsive gene networks we identified in Chapter 3. Several components of the GABA-A receptor were also among this list, including the α2, β2 and β3 subunits. Ethanol is a well known allosteric modulator of GABA A receptors (Nestoros, 1980) and acute ethanol exposure increases GABA activity in the short term (Breese et al., 2006), and GABA-A receptors that contain the β3 subunit show a much stronger sensitivity to lower doses of acute ethanol (Wallner et al., 2003). As mentioned in section 1.2.1, polymorphisms in the GABA α2 subunit gene, GABA A receptor, subunit α2 (*Gabrb2*), are associated with lifetime alcoholism (Edenberg et al., 2004) and may impact the rewarding effect of alcohol consumption (Pierucci-Lagha et al., 2005).

The data mentioned in section 2.5, concerning the large overlap between expression patterns derived here for acute ethanol and a published study on basal gene expression correlating with predisposition to ethanol consumption (Mulligan et al., 2006), does lend strong support to the argument that the ethanol responsive genes discussed here are largely related to direct ethanol actions. Together, these overlaps between the current ethanol responsive datasets and other genomic data, strongly suggest that direct actions of ethanol were the major factor deriving the ethanol responsive genes identified in this work.

This genetic analysis of ethanol-responsive gene expression allowed extension beyond dichotomous gene lists, to the spectrum of acute ethanol transcriptional responses influenced by naturally occurring polymorphisms segregating in the BXD strains. This approach identified gene groups characterized by a wide range of differential expression profiles: including genes such as *Npas4*, which was consistently up-regulated by ethanol, and *Gsk3β*, whose response entailed up-regulation, down-regulation and no change, depending on the subset of strains Figure 2.6. Such a range of expression changes not only highlights the complex role of genetic background in modifying molecular responses to acute ethanol exposure, but also suggests these genes play a role in mediating behavioral responses to acute ethanol, which manifest in a similarly divergent manner. In Chapter 3 we extend this work to identify specific behavioral phenotypes that may be directly linked to some of the genetic responses to acute ethanol described here.

2.6.1 Limitations

One potential confound in our analysis of ethanol-responsive gene networks regards the experimental design used for the microarray studies. Since the BXD strains used for tissue harvesting were also part of a behavioral genetics analysis on ethanol anxiolytic actions, the animals received mild restraint stress and behavioral testing in addition to saline or ethanol treatment before harvesting tissues 4 hours after drug treatment (as described in section 2.3.1). Our analysis using S-scores to compare saline versus ethanol-treated animals was calculated to at least partially remove the effect of stress from the derived expression patterns, since both groups were handled identically. Still, we cannot rule out that an interaction between stress and ethanol, rather than just a response to acute ethanol, might contribute to some of the transcriptional responses

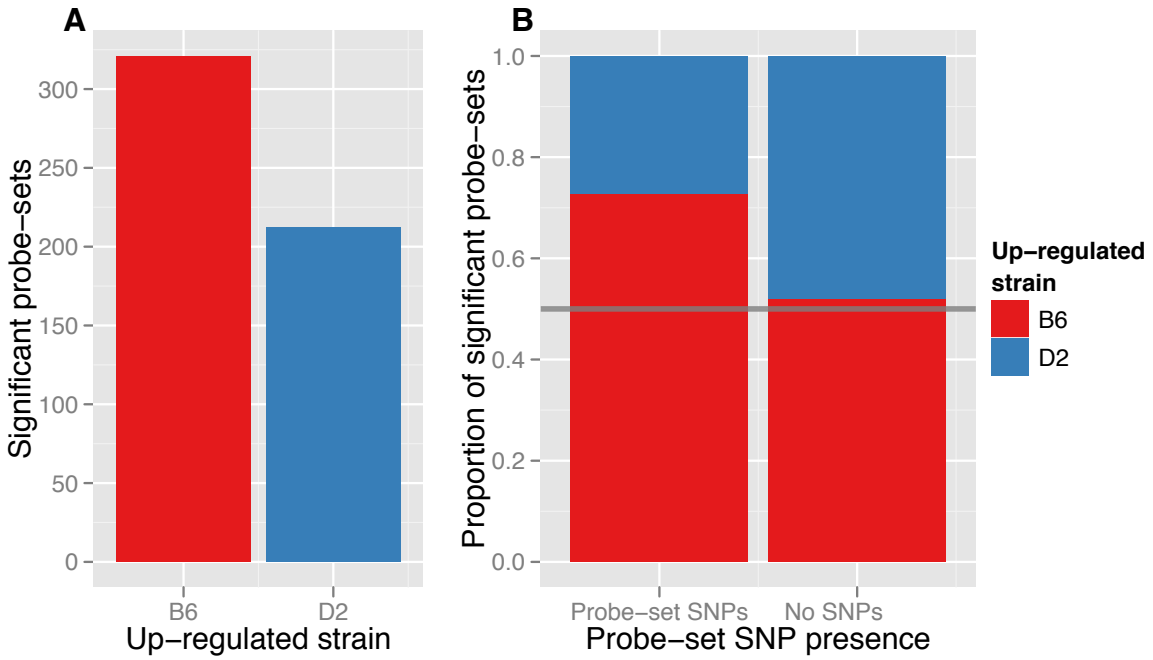

Figure 2.13. Impact of polymorphic probe targets on differential expression.

reported here.

2.6.2 B6/D2 polymorphisms overlapping microarray probes

An important consideration when working with microarray data is the potential impact of polymorphisms on probe/target hybridization. In situations where a probe's intended target harbors a polymorphism, disrupting their prefect complementarity, the reported measurement of transcript will be tainted by altered hybridization kinetics (Gilad et al., 2005). This is especially important in the context of eQTL studies, which were performed in Chapter 3, because allele specific hybridization artifacts can manifest as spurious *cis* eQTLs, inflating the number of false positives. Doss et al. (2005) found that only 10 out of 28 putative *cis* eQTLs mapped using an B6 × D2 F_2 intercross could be validated in molecular follow-up experiments.

Since much of this project relied on microarray expression data from animals that

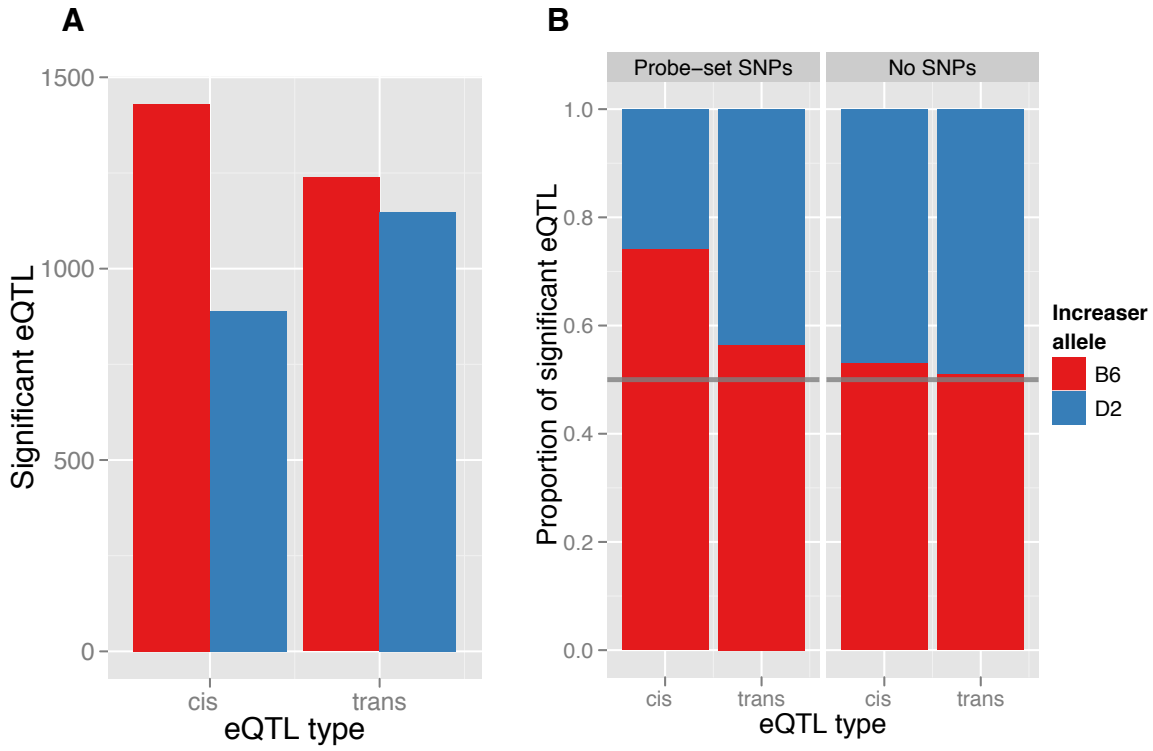

Figure 2.14. Impact of polymorphic probe targets on eQTL mapping.

harbor both B6 and D2 alleles, I wanted to mitigate the effects of polymorphic probe targets as much as possible. Using the D2 genome sequence provided by Dr. Rob Williams, I attempted to identify all probes on the M430v2 GeneChip that may be affected by the presence of D2 SNPs. Because this analysis requires repeating when working analyzing data from different genetic samples or using different microarray platforms, I wrote a set of R functions that attempts to take much of the pain out of this process, which are provided in appendix A.4.

Using these functions, I identified 14,247 M430v2 probes that overlap at least one D2 SNP. These correspond to 6,439 individual probe-sets, which means ≈ 12% of all probe-sets are potentially affected. The full list of affected probe-sets is provided in Table S9. To gain an appreciation of how these polymorphic probe targets could impact our data, I performed a simple differential expression analysis between B6 and D2 PFC

samples; 533 probe-sets were significant at a FDR adjusted p-value of 0.05. Figure 2.13A reveals that a disproportionately high number of these probe-sets were up-regulated in B6. This bias was almost entirely abrogated upon the removal all SNP-affected probe-sets (Figure 2.13B). I also examined the impact of these on eQTL mapping results. Dividing all significant eQTL in the PFC saline data-set into groups based on which genotype is the increaser allele produced a similar bias. As displayed in Figure 2.14, for a disproptintely large number of *cis* eQTL, B6 was the increaser allele. And again, this bias was almost entirely wiped out by removing SNP-affected probe-sets (Figure 2.14B). These analyses clearly demonstrate the potentially major impact SNP-affected probes can have on the results of genomic analyses. As such, we used the list provided in Table S9 to filter out affected probe-sets from all relevant analyses, especially those relying on the presence of *cis* eQTLs to prioritize positional candidate genes.

Chapter 3

Genetic analysis of ethanol responsive networks

In order to extract and dissect acute ethanol-responsive gene networks, we performed a large-scale gene expression analysis across RI strains derived from the BXD genetic mapping panel. The BXD family has been widely used for both genetic studies on ethanol behaviors and many other phenotypes, and for expression genetics studies (Chesler et al., 2005). For each included BXD strain, we utilized PFC, VMB and NAc microarray expression profiles from saline and ethanol treatment groups that were obtained as part of a previous project conducted in the Miles laboratory (Putman, 2008). This produced the most robust assessment of ethanol-responsive brain gene expression to date. Furthermore, we focused on PFC and produced the first genetic analysis of ethanol-responsive gene networks. Our results show network-level enrichment of genes involved in synaptic plasticity and identify key hub genes regulating the ethanol response for large networks of genes. This first such detailed genetic analysis of the acute ethanol responsome may provide valuable insight for molecular mechanisms underlying the neurobiology of ethanol and also ultimately provide novel AUD susceptibility candidate genes and targets for intervention in alcoholism.

3.1 Constructing gene co-expression networks

As we alluded in section 1.4, various methods exist for generating gene co-expression networks. The simplest method involves calculating Pearson correlations for all pair-wise genes and applying a hard threshold to determine which genes should be connected. The robustness of these networks, initially called *relevance networks*, can be then assessed through permutation testing (Butte and Kohane, 2000). However, the reliance on hard thresholds to classify the relationship between genes as either connected or unconnected is a potential limitation of relevance networks. As the dichotomy imposed by this approach may artificially limiting and cause biologically meaningful relationships to be overlooked (Carter et al., 2004). More rigorous approaches for constructing gene co-expression networks avoid this potential pitfall by utilizing "soft-thresholds."

3.1.1 WGCNA

One such method is the increasingly popular weighted correlation network analysis (WGCNA) approach first described by Zhang and Horvath (2005). WGCNA uses soft-thresholding to generate networks that conform to a scale-free topology. Scale-free networks follow the power distribution they are named for, comprising many nodes that have sparse connections and a few that are highly interconnected. In addition to providing an accurate model for metabolic networks (Jeong et al., 2000), neural networks of the roundworm *C. elegans* (Watts and Strogatz, 1998), and the World Wide Web (Barabasi and Albert, 1999), the scale-free topology also typifies gene co-expression networks (van Noort et al., 2004). Some researchers recently have used WGCNA to define correlated gene modules associated with blood alcohol levels using the "drinking-in-the-dark" paradigm of excessive ethanol consumption in B6 mice (Mulligan et al.,

2011). WGCNA is implemented as a freely available package for R (Langfelder and Horvath, 2008).

3.1.2 Paraclique analysis

Paraclique analysis is another example of a rigorous approach to constructing gene co-expression networks that utilizes soft-thresholding (Baldwin et al., 2005). Unlike WGCNA, this approach uses a graph theoretical algorithm to identify gene co-expression networks within a given data-set (Chesler and Langston, 2005). The most natural grouping of vertices in a graph is by cliques, or fully connected subgraphs. While finding the maximum clique is a well-known computationally intractable problem, the topology of biological graphs lends itself to solution by advanced algorithmic implementations (Abu-Khzam et al., 2006; Langston et al., 2008). The inevitable noise in large microarray datasets can render cliques too restrictive, as a single missing correlation between two genes would abrogate the entire clique.

Paracliques are the relaxed version of a clique, which makes allowances for missing edges within a graph and are therefore well suited for the analysis gene expression microarray data. For graphs constructed using a correlation threshold, maximum cliques are iteratively extracted and used as cores on which to build paracliques. A paraclique starts with a maximum clique and gloms onto all vertices with at least some proportion of edges to that clique. This proportion is called the *proportional glom factor*. As a paraclique is formed, the number of edges that must be present for a vertex to be included is scaled to the size of the starting clique. Unlike WGCNA, paracliques make no assumptions about the topology of generated networks.

We utilized the paraclique approach to conduct the network analyses described in this thesis. An empirical assessment of WGCNA, paraclique analysis and other clustering

methods found that paraclique-generated gene modules were consistently more enriched for known biological pathways across small, medium and large gene sets (Eblen et al., 2011). Still, limited testing of both paracliques and WGCNA with our microarray expression data-sets produced results that were largely overlapping.

3.2 Gene network analysis in prefrontal cortex

Rather than focusing on gene-lists, as was only possible in the analysis of B6 and D2 strains previously published by our lab (Kerns et al., 2005), we used the power of genetic correlations across the BXD strains to derive coherent gene networks. Due to the complexity of this analysis and the importance of the PFC in influencing long-term adaptive responses to ethanol and goal-directed behavior (Kalivas et al., 2005; Liu et al., 2006; Robinson and Kolb, 1997), we restricted our network analysis to this brain region.

3.2.1 Paraclique construction

Steady-state RMA and saline versus ethanol S-score expression datasets were analyzed using the paraclique-finding algorithms described above. We first calculated all pairwise Pearson correlations across probe-sets, where each probe-set is represented as a vector of BXD expression values, and used this data to construct an unweighted graph in which vertices represent probe-sets and edges were present whenever the absolute value of the correlation between two probe-sets was ≥ 0.7. The choice of threshold when converting a weighted graph to an unweighted graph is analogous to the choice of p-value when determining significance; it is chosen to produce a reasonable tradeoff between false positives and false negatives. A correlation threshold of $|0.7|$ across 27 strains yields a correlation p-value of 4.8×10^{-5} (calculated using Student's t-distribution). Such low p-values are indicative of the rigor of graph theoretical techniques.

We selected a glom factor of 0.7 for the analyses presented here, which maintains an edge density > 90% in nearly all the resulting paracliques. For such defined paracliques, probe-sets had expression responses to ethanol correlated with at least 70% of the other paraclique members at a threshold $\geq |0.7|$. Lowering the glom factor below 0.7 resulted in a sharp drop-off in edge density. Furthermore, empiric testing showed that more stringent glom factors produced similar overall functional results but tended to fragment known correlated gene groups (e.g. dopamine signaling genes) into multiple paracliques.

3.2.2 Network topology

The relative importance of each node within a paraclique was assessed using network topological measures of connectivity and centrality. Degree of connectivity was equal to number of edges linking a probe-set to other paraclique members, based on the |0.7| edge correlation threshold used to construct the unweighted graphs. Betweenness centrality measures how frequently a node is included in the shortest paths between all pair-wise members of a network. With the edge threshold at |0.7|, Spearman's rank correlations were typically > 0.9 between centrality and connectivity. Increasing the edge correlational threshold to |0.9| reduced the connectivity/centrality correspondence to ≈ 0.6 and greatly increased the centrality for a subset of nodes situated between densely inter-connected subnetworks. We therefore used betweenness centrality scores within unweighted graphs constructed using the more stringent |0.9| edge threshold as a supplemental measure of node importance. Both measures were calculated using the `igraph` package for R (Csardi and Nepusz, 2006).

Fisher's exact test was used to identify paracliques that harbored a greater number of significantly ethanol-responsive probe-sets than what would be expected by chance.

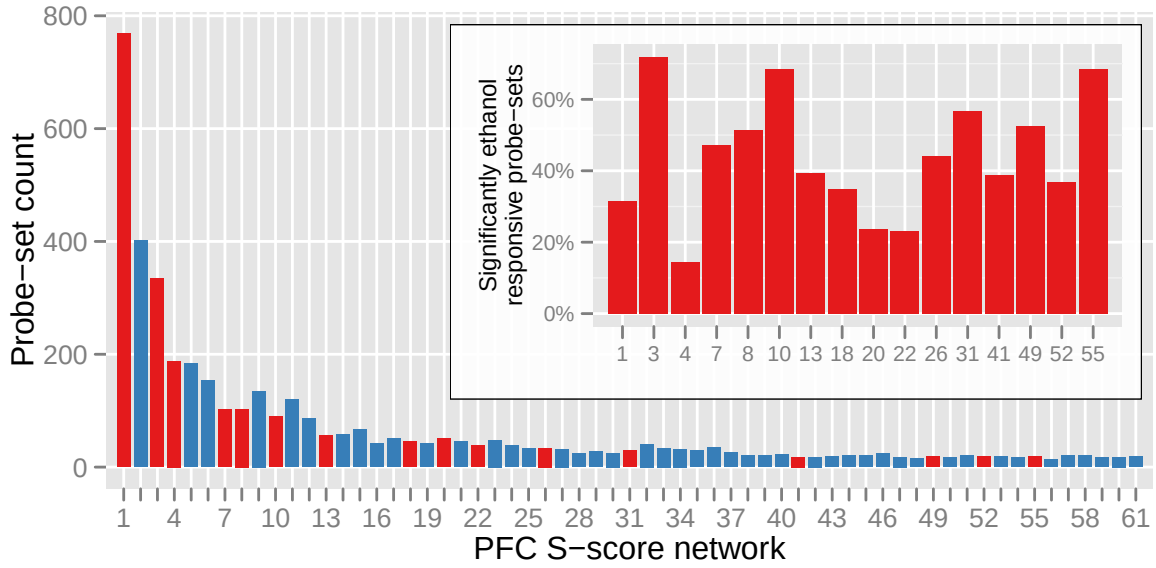

Figure 3.1. PFC saline versus ethanol S-score paraclique networks. Distribution of S-score network sizes based on the number of genes assigned to each. Significantly ethanol-responsive genes were over-represented in a subset of these networks (red bars). These 16 paracliques, shown in the inset, were considered ethanol responsive gene enriched networks (ErGeNs).

The 30,941 probe-sets that passed the present-call filter (section 2.3.4) served as the background for this analysis. Paracliques with a Bonferroni adjusted p-value ≤ 0.05 were judged to be significantly enriched for ethanol-responsive probe-sets.

3.2.3 Saline versus ethanol S-score paraclique networks

Using saline versus ethanol S-score data, a total of 61 paraclique networks were identified in the PFC, while 118 and 93 paracliques were identified in the saline and ethanol RMA data-sets, respectively. The full lists of probe-sets that constitute each paraclique network are provided in Table S4. Each paraclique represents a densely intercorrelated groups of genes. In the saline and ethanol networks formed with RMA expression datasets, inter-gene correlations represented the admixture of treatment variation superimposed on basal steady-state mRNA levels. In the context of the S-score networks, the

correlations strictly reflect coordinated changes in expression induced by acute ethanol. The size of the S-score paracliques ranged from 710 to 11 probe-sets (Figure 3.1). While 64% of all significantly ethanol-responsive genes in the PFC belonged to one of the 61 S-score networks, a Fisher's exact test revealed a subset of networks that were statistically enriched for these genes. These ethanol responsive gene enriched networks (ErGeNs) are depicted in the inset of Figure 3.1.

Network-based clustering (Figure 3.3) and a traditional non-parametric cluster analysis (Figure 3.2) of all significantly ethanol responsive genes, both revealed the existence of several modules of co-expressed genes that were largely subcomponents of these paraclique-derived ErGeNs, most predominantly ErGeN1 and ErGeN3. Taken together, these results suggested that, at the time point employed by these studies, the PFC transcriptional response to acute ethanol was primarily mediated through a relatively small number of highly organized gene networks.

3.2.4 Cross-treatment network comparisons

To determine how networks generated from the different treatment groups (saline versus ethanol RMA networks) and analyses (saline/ethanol RMA networks vs S-score networks) related to each other, we performed pairwise comparisons of all network members (Table S5).

Many of the saline networks significantly overlapped with networks in the ethanol data, indicating the inter-gene correlations that constitute these networks are largely stable across treatments and likely represent robust biological relationships. Similarly, S-score networks generally had a substantial and predominant relationship with a single or small number of saline or ethanol networks, as would be expected given that S-scores were derived from the same data-sets.

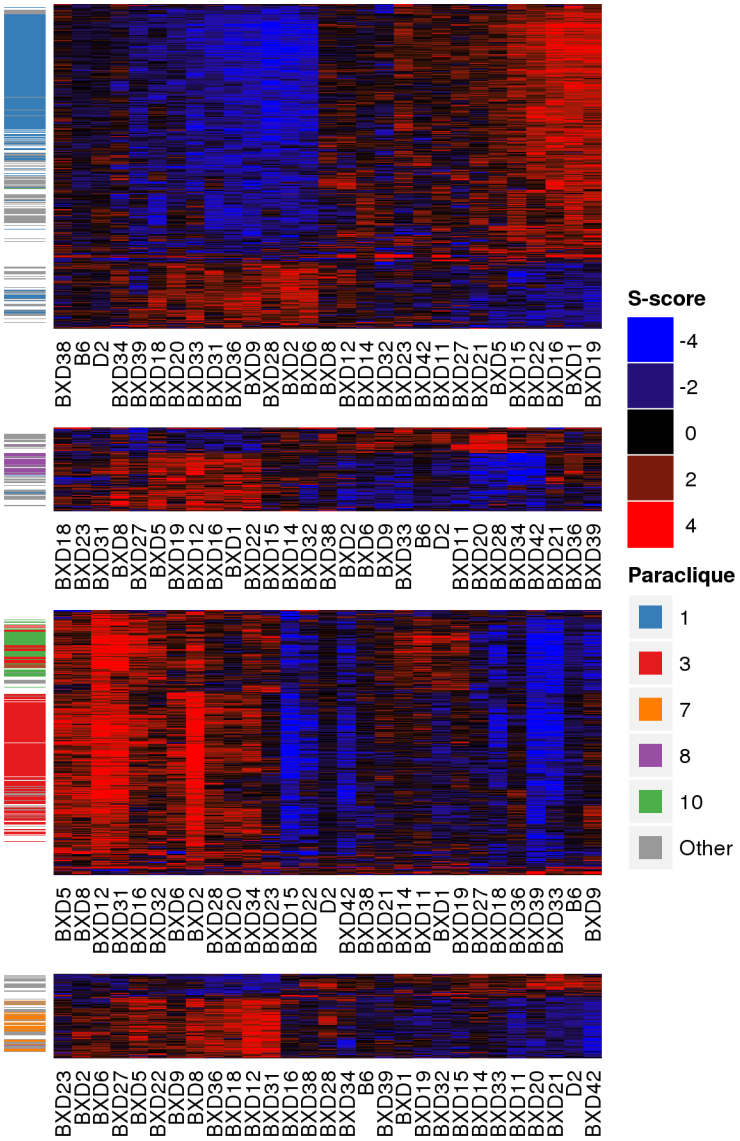

Figure 3.2. Non-parametric cluster analysis of all significantly ethanol-responsive genes in the PFC. Genes were assigned to modules using *k*-means clustering. The number of modules was determined by principal component analysis, which revealed the first 4 components explained ≈70% of the variation in the S-scores for these genes. Each module was hierarchically clustered independently based on average linkage of Pearson correlation distance. These results are visualized in the above heatmap, where warmer colors indicate positive S-scores (up-regulated by ethanol) and cooler colors indicate negative S-scores (down-regulated by ethanol). The adjacent column of colors indicates to which S-score paraclique network the corresponding gene was assigned in the PFC.

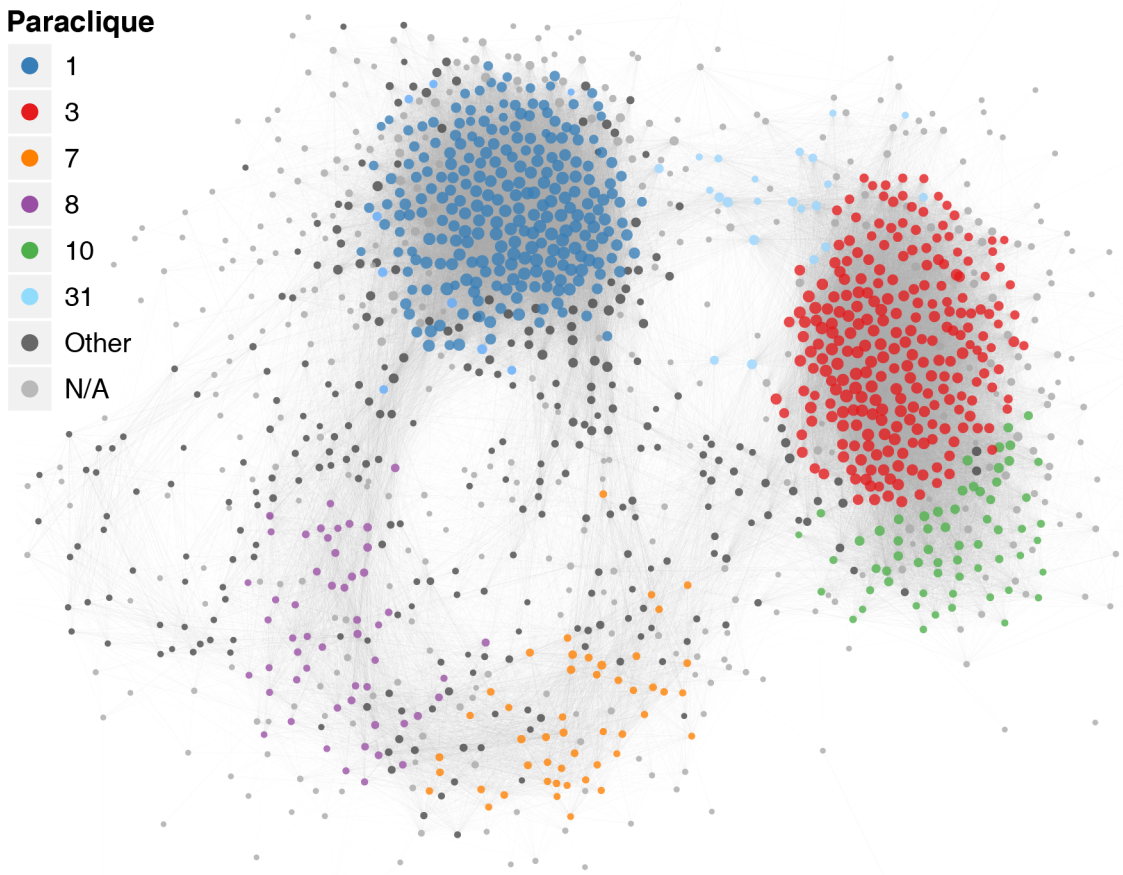

Figure 3.3. Network-based clustering of ethanol responsive genes in the PFC. Network-based clustering of the 1,246 significantly ethanol responsive genes in the PFC revealed distinct modules largely corresponding to the ErGeNs depicted in Figure 3.1

We examined in detail how the two major networks comprising the PFC transcriptional response to ethanol, ErGeN1 and ErGeN3, related to their respective counterparts in the saline and ethanol expression data, in order to determine what additional information is provided by the S-score networks. ErGeN1 was significantly enriched for members of saline network 1 and ethanol network 1. Likewise, the gene members of saline network 1 and ethanol network 1 significantly overlapped each other, with 215 genes in common. The overlapping components of these three networks were frequently the mostly highly connected nodes (Figure 3.4, ErGeN1 panel). ErGeN3 exhibited a similar relationship with saline network 4 and ethanol network (Figure 3.4, ErGeN3 panel). Therefore, these S-score networks largely comprise the robustly inter-connected hubs of existing networks. However, missing from Figure 3.4 are the 439 and 143 probe-sets that belong to ErGeN1 and ErGeN3, respectively, but not their counterpart networks in the saline or ethanol RMA expression data. These network facets unique to the ErGeNs represent a form of genetic co-regulation that would have gone undetected without the use of S-score data.

3.3 Genetic regulation of ethanol-responsive networks

3.3.1 eQTL mapping

To uncover the genetic elements regulating these networks we performed eQTL mapping for each probe-set's expression trait in the saline RMA and S-score data-sets. This analysis was performed using a subset of informative microsatellite and SNP markers that have been used to genotype the BXD family (Shifman et al., 2006; Williams et al., 2001), and are available from GeneNetwork at `http://genenetwork.org/genotypes/BXD.geno`. Linkage between genotypes and expression phenotypes was

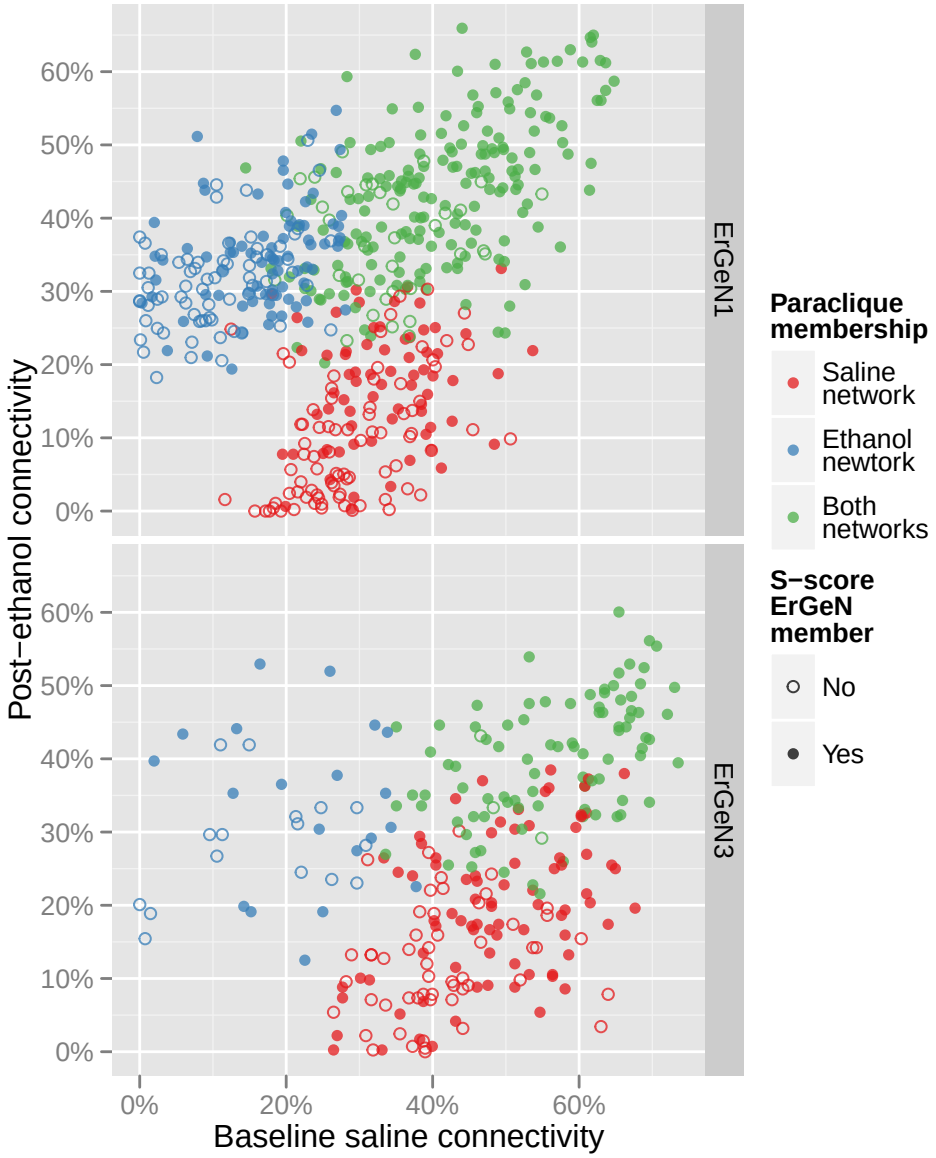

Figure 3.4. Cross-treatment network connectivity. Relationship between ErGeNs and counterpart networks in RMA expression data. Both S-score networks, ErGeN1 and ErGeN2, had counterpart networks in the basal saline and post-ethanol expression data: ErGeN1 significantly overlapped with saline network 1 and ethanol network 1; ErGeN3 significantly overlapped with saline network 4 and ethanol network 2 (Table S5). Each point represents a gene that belongs to a given ErGeN's counterpart saline network (blue), ethanol network (red) or both (green). Filled-in points indicate the gene also belongs to the overlapping ErGeN. The x- and y-axes measure gene connectivity ($|$Pearson correlation coefficient$| \geq 0.7$) within the saline and ethanol expression datasets, respectively.

assessed by Haley-Knott regression using `R/qtl` (Broman et al., 2003; Haley and Knott, 1992).

We corrected for multiple testing and obtained significance thresholds using a permutation analysis approach, where phenotype/genotype associations were recalculated multiple times using randomly shuffled versions of the observed phenotype data (Doerge and Churchill, 1996); after each permutation, the maximum logarithm of odds (LOD) score was recorded. For each probe-set expression trait, this process was repeated 1,000 times. Genome-wide adjusted p-values were then obtained by calculating the percentage of LOD scores from the randomly permuted data that exceeded a given trait's observed LOD score.

We classified the significance of an eQTL using guidelines put forth by the Complex Trait Consortium for mapping traditional eQTL (Abiola et al., 2003) where 'significant' refers to genome-wide corrected p-values ≤ 0.01 and 'suggestive' refers to p-values ≤ 0.63. Estimates of true QTL location were obtained using `R/qtl` to calculate 1.5 LOD score drops, as recommended by Manichaikul et al. (2006). eQTL were considered *cis* eQTL if their peak chromosomal location was less than 5 Mb upstream or downstream of the regulated gene, all others were considered *trans* eQTL.

3.3.2 *Trans*-band analysis

Loci enriched for *trans* eQTL, referred to as *trans*-bands, were detected by splitting the genome into 10 Mb bins and counting the number of suggestive eQTL that mapped to each. In order to determine whether a particular genome bin harbored more eQTL than would be expected by chance, we performed 10,000 permutations, each involving random assignment of all eQTL to a genetic marker and recording the number of mappings at the most populous bin. Observed *trans*-bands were deemed significant if

they exceeded the 95th percentile of the distribution of peak *trans*-bands captured from each permutation. To facilitate the search for candidate regulators underlying these eQTL enriched regions, we defined support intervals for each of the major *trans*-bands by aggregating the support intervals calculated for the individual eQTL comprising each *trans*-bands. *Trans*-bands support intervals were defined as the chromosomal regions flanked by genetic markers that were included in at least 80% of the *trans*-band member's individual support intervals. Figure 3.9 provides a visualization of several ErGeN *trans*-bands support intervals to clarify this approach.

3.3.3 Saline and S-score eQTL profiles

Performing eQTL mapping across both the saline and S-score data-sets allowed us to assess how the baseline regulatory architecture of the PFC transcriptome is altered by exposure to acute ethanol. The genetic regulatory profiles for the RMA and S-score datasets differed substantially. Although the majority of probe-sets mapped to at least one suggestive eQTL (Table 3.1), only 6% of eQTL positions were conserved in both the saline and S-score datasets. Indeed, we observed a fundamental shift in the type of genetic regulation most prominent across these datasets. Of the 3,279 genes with significant eQTL in the saline expression data, 42% were considered to be *cis*-acting, since the peak eQTL location mapped within 5 Mb of the linked expression trait. Whereas in the S-score data *cis* eQTL accounted for less than 1% of the 1,215 genes with significant eQTL. The full eQTL mapping results are provided in Table S6.

The effective absence of *cis* eQTL in the S-score data suggests that mechanisms underlying ethanol-responsive gene regulation may fundamentally differ from those governing basal transcription. However, some portion of the basal *cis* eQTL are likely spurious associations driven by polymorphisms between the B6 and D2 genomes that

Table 3.1. Expression QTL mapping results for saline RMA and S-score data sets

Data set	eQTL class	Suggestive eQTL (p-value < 0.63)	Significant eQTL (p-value < 0.05)
Saline	*trans*	9,570	1,877
	cis	433	1,355
S-scores	*trans*	10,968	1,276
	cis	62	7

Suggestive and significant p-value thresholds are genome-wide corrected.

affect microarray probe target hybridization (Alberts et al., 2007; Doss et al., 2005). As the impact of such SNP effects should be invariant across the saline versus ethanol treatment conditions, any spurious *cis* eQTL would be effectively filtered out of the S-score eQTL results.

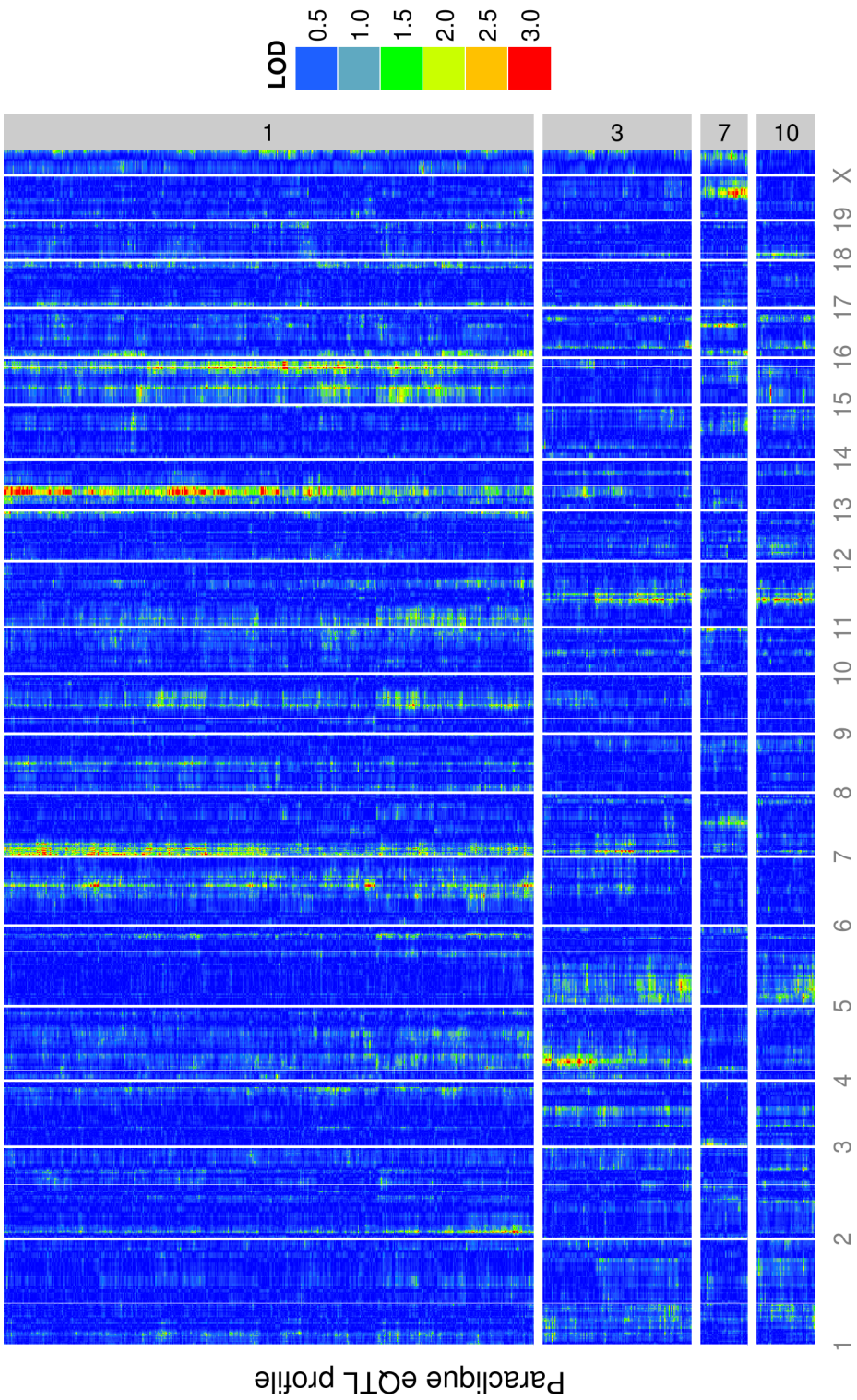

Figure 3.5. Ethanol responsive network eQTL profiles Cluster analysis of eQTL profiles for all genes that belong to either ErGeN1, ErGeN3, ErGeN7 or ErGeN10 and have at least one suggestive S-score eQTL (genome-wide corrected $p-value < 0.63$). The eQTL heatmap indicates the strength of association between a probe-set's S-scores (rows) and each genetic marker (columns) across the genome, warmer colors indicate stronger linkage.

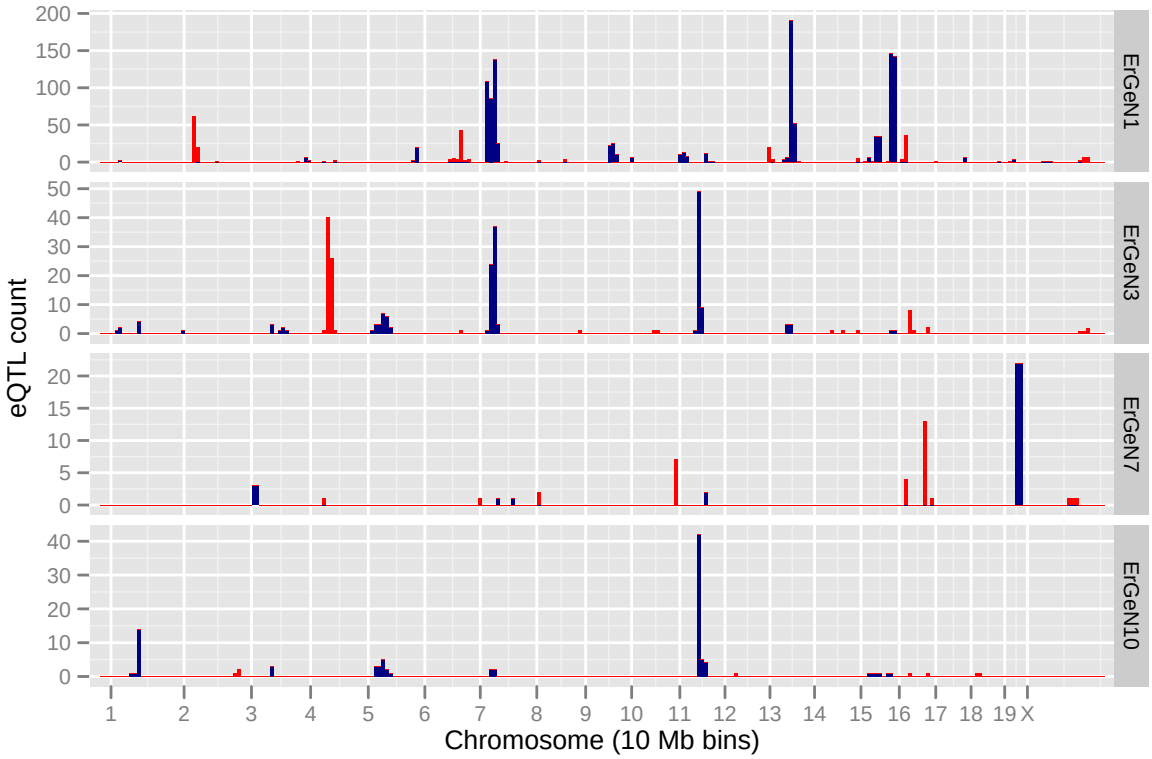

Figure 3.6. Ethanol responsive network *trans*-bands Histogram of S-score eQTL (genome-wide p-value < 0.63) frequencies across the genome divided into 10 Mb bins. This representation of the eQTL data faciliates the identication of *trans*-bands. For example, many members of ErGeN3 are linked to the proximal region of Chr 7, which is obscured in Figure 3.5.

Similar to other genetical genomics studies, we found that many changes in transcript abundance induced by acute ethanol were linked to a relatively small number of highly influential loci, so-called *regulatory hotspots* or *trans*-bands. This was particularly salient for eQTL profiles of the major ErGeNs (Figure 3.5). As shown in Figure 3.6, these networks could largely be partitioned into 6 *trans*-bands that mapped to loci on Chrs 4, 7, 11, 13, 15 and 19. In most cases, these *trans*-bands were unique to specific networks, the exceptions being the Chr 7 and Chr 11 *trans*-bands, which were composed of genes from ErGeN1 & ErGeN3, and ErGeN3 & ErGeN10, respectively (Table 3.2).

Table 3.2. ErGeN trans-band support intervals

Chr	ErGeN ID	Peak Mb	Peak marker	Support interval
4	ErGeN3	46.61	rs13477694	35.49–55.07
7	ErGeN1	34.62	rs3694031	15.52–36.48
7	ErGeN3	30.14	rs8261994	24.06–30.43
11	ErGeN3	58.38	rs3697686	53.89–68.93
11	ErGeN10	58.38	rs3697686	56.35-62.07
13	ErGeN1	54.88	rs13481817	47.68–69.04
15	ErGeN1	89.87	rs13482702	86.80–95.78
19	ErGeN7	41.69	rs3653396	32.73–41.95

3.4 Hub genes within ethanol-responsive networks

The parameters used to construct the networks described above were such that the vast majority of genes share edges with at least half of the remaining network. Subsets of genes shared edges with nearly all network members, and were more important to the network based on measurements of connectivity and centrality. These network hub genes could be major regulators of the transcriptional response to acute ethanol and more generally, may represent key points of vulnerability in underlying signaling pathways responding to ethanol. We therefore identified hub genes by ranking network members based on their degree of connectivity and betweenness centrality (Table S4).

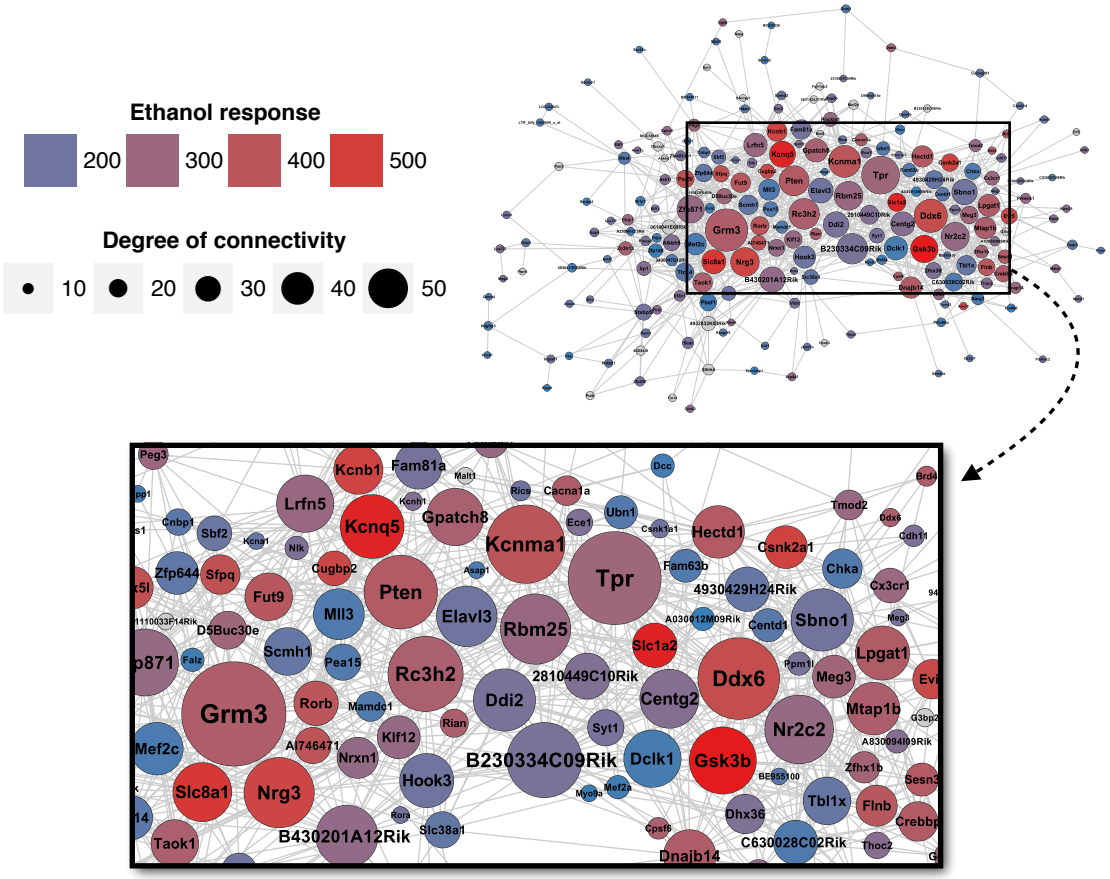

Figure 3.7. Ethanol responsive gene-enriched network 3 Network visualization of all genes comprising ErGeN3 that share at least one adjacent edge at a correlation threshold of $> |0.9|$. Node color indicates the magnitude of a gene's transcriptional response to ethanol, quantified using Fisher's combined p-values. Grey nodes were unaltered by ethanol. Node size represents a genes degree of connectivity.

Table 3.3. Candidate genes within ErGeN *trans*-band support intervals

Trans-band	Gene	ErGeN	Diff. exp. q-value[a]	Scaled connectivity	Scaled centrality	*cis* eQTL p-value: saline	*cis* eQTL p-value: S-score	Coding SNPs	Non-syn SNPs[b]
Chr 7	*Aplp1*	1	0.04	0.95	0.92	0.004		6	3
Chr 7	*Scn1b*	1	0.02	0.94	0.75	0.05	0.51	6	1
Chr 11	*Gria1*	3	0.03	0.49	0.56	2e10-5		16	0
Chr 11	*Ncor1*	10	0.005	0.03	0.75		0.19	14	5
Chr 13	*Sncb*	1	0.008	0.98	0.92	0.2	0.21	1	0
Chr 15	*Nell2*	1	0.002	0.94	0.99	0.01		4	2

[a] Differential expression q-values described in section 2.4.
[b] Non-synonymous SNPs.

Among the most highly connected hubs within ErGeN3 were a number of genes that have been previously implicated in modulating level of response to ethanol or susceptibility to alcohol dependence (Figure 3.7), including *Kcnma1* and *Gsk3β*. *Kcnma1* is a large conductance potassium channel whose activity is directly affected by ethanol (Dopico et al., 1996). *Gsk3β*, is a serine/threonine kinase that participates in the WNT signaling pathway and is an important modulator of ethanol-induced neurotoxicity in both mice (Chen et al., 2009a) and Drosophila (French and Heberlein, 2009). Our own recent work has shown that over-expression of *Gsk3β* in mouse PFC alters ethanol consumption (Meng et al., manuscript submitted). These findings on *Kcnma1* and *Gsk3β* serve to validate our network analysis approach, identifying these and other hub genes (Figure 3.7) as potentially important modulators of ethanol phenotypes.

The ErGeN3 member with the highest degree of connectivity was a probe-set (`1435583_at`) unhelpfully annotated as AU067633. However, recent data from high-throughput RNA sequencing (RNA-seq) analysis of B6 and D2 brain transcripts (Lu and Williams, personal communication) strongly suggests that this probe-set actually targets the distal 3′ untranslated region of *Grm3*, a metabotropic glutamate receptor (Figure 3.8). Given the considerable evidence that metabotropic glutamate receptors are key mediators of the neuroadaptations associated with addiction (Gass and Olive, 2008), *Grm3*'s position as a major hub of this ethanol-responsive network has mechanistic implications for regulation of the network and further supports the overall significance of this network in ethanol traits.

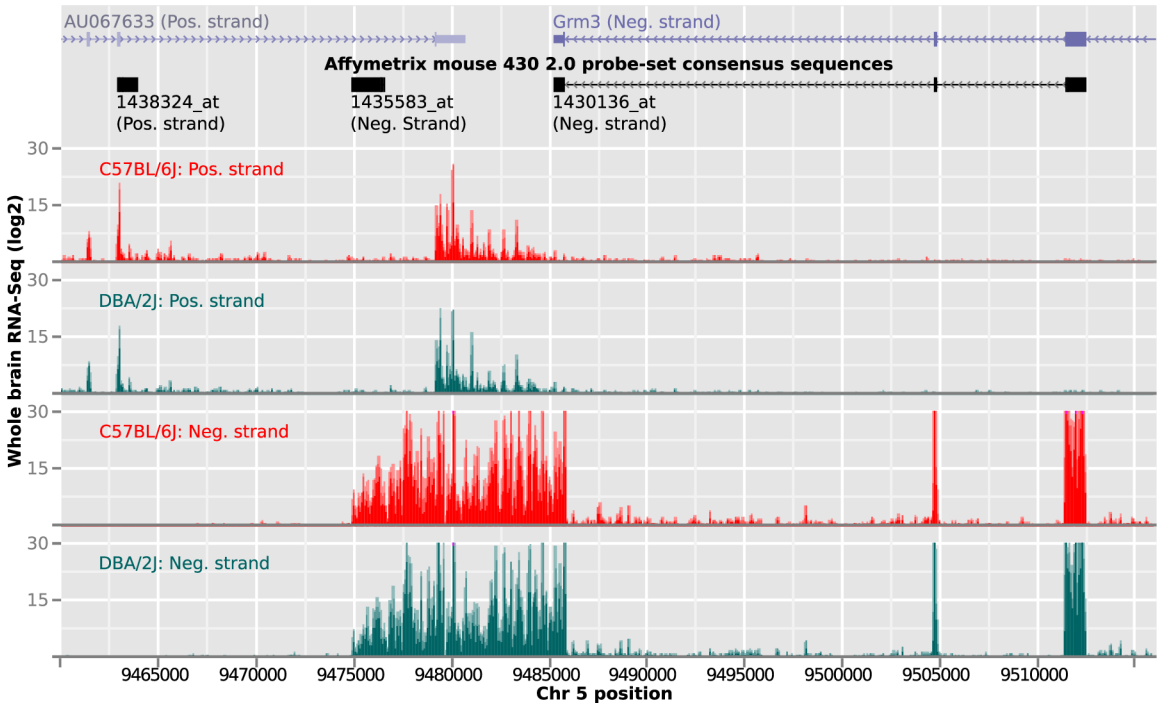

Figure 3.8. Whole brain RNA-seq expression data across the Chr 5 region that encompasses AU067633 and *Grm3*, adapted from the GeneNetwork mirror of the UCSC Genome Browser (ucscbrowser.genenetwork.org). Although probe-set 1435583_at (red) putatively maps to an AU067633 intron, it appears to actually target *Grm3*'s 3′ UTR, which is highly expressed from the negative strand across the same stretch of DNA. Probe-set 1435583_at's basal RMA expression levels were significantly correlated with the distal *Grm3* probe-set, 1430136_at ($r = 0.77$, p-value = 2.2E-06), while showing no relationship to the proximal AU067633 probe-set, 14338324_at ($r = 0.29$, p-value = 0.14).

3.5 Candidate regulators of ethanol-responsive networks

3.5.1 Prioritizing positional candidate genes

Candidate regulators for *trans*-bands were derived by an empiric ranking scheme for genes located within the support interval of the *trans*-band. This ranking scheme assigned points for gene information within four categories: genetic sequence variation (SNPs), expression genetics (*cis* eQTL), ethanol regulation and network properties. Positional candidates were scored based on harboring polymorphisms between the B6 and D2 genomes that may alter protein function. Genes carrying non-synonymous or functional polymorphisms were considered higher priority candidates. We also took into account non-coding polymorphisms whose functional impact may only manifest at the transcript level by further prioritizing interval candidate genes associated with a robust *cis* eQTL in either the saline RMA or S-score expression datasets. In order to prevent false positive *cis* eQTLs from being prioritized, probe-sets with *cis* eQTL were penalized if their binding target region contained a D2 polymorphism identified as part of the analysis described in section 2.6.2. As Affymetrix probe sequences were designed against the B6 genome, probe SNPs should only reduce binding avidity with D2 transcripts. Therefore, this penalty was only applied to *cis* eQTL if B6 was the increaser allele. Candidates were prioritized further if they belonged to the same network as constituents of the linked trans-band, taking into account the relative importance of a gene in the resident network by using the connectivity and centrality measures from the hub gene analysis. Genes identified as significantly ethanol-responsive across the BXD lines received additional scoring. The full list of ranked candidate genes for each *trans*-band is provided in Table S7.

The two largest ethanol-responsive networks, ErGeN1 and ErGeN3, shared a common

regulator on the proximal end of Chr 7, between 15.52 and 36.48 Mb (Table 3.2). Examination of eQTL for all members of these networks revealed a complicated pattern of association, in which the *trans*-band could be subdivided into several groups based on peak eQTL locations that clustered between 16.3 and 35.04 Mb (Figure 3.9). Peak linkage of genes from ErGeN3, however, was limited to a narrow region between 30.1 and 30.2 Mb, at the distal edge of the support interval. This locus represents the common regulatory hot-spot shared by these two networks and harbors the two most highly ranked candidate regulators of the Chr 7 *trans*-band: *Scn1b*, a voltage gated sodium channel subunit and *Aplp1*, amyloid beta precursor-like protein (Table 3.3). Both genes were significantly ethanol-responsive, highly connected hub nodes in ErGeN1 and associated with *cis* eQTL in the saline data. Unlike *Aplp1*, the ethanol response of *Scn1b* was at least partially regulated by a local polymorphism, as evidenced by its suggestive *cis* eQTL in the S-score data. Both genes contain coding polymorphisms, *Aplp1*'s harbored a polymorphic splice site, raising the possibility that different *Aplp1* isoforms may segregate members of the BXD family.

Of the ErGeN1 genes without a *trans* eQTL on proximal Chr 7, most could be partitioned into *trans*-bands linked to Chr 13 or 15. The regulatory hotspots underlying these *trans*-bands were both unique to ErGeN1 (Table 3.2). The Chr 13 *trans*-band support interval spanned from 47.6 to 69 Mb and peaked at 54.88 Mb. QTL for both cocaine induced activation (Gill and Boyle, 2003) and hypothalamic CRF binding protein (*Crf-bp*) transcript abundance (Garlow et al., 2005) were previously mapped to this region. Ranking the positional candidates within this region revealed a promising candidate in synuclein β (*Sncb*), a neuronal protein that is widely co-localized to presynaptic terminals throughout the brain (Chandra et al., 2004). *Sncb* was one of the largest ErGeN1 hub genes and was regulated by suggestive *cis* eQTL in both the saline and S-score datasets.

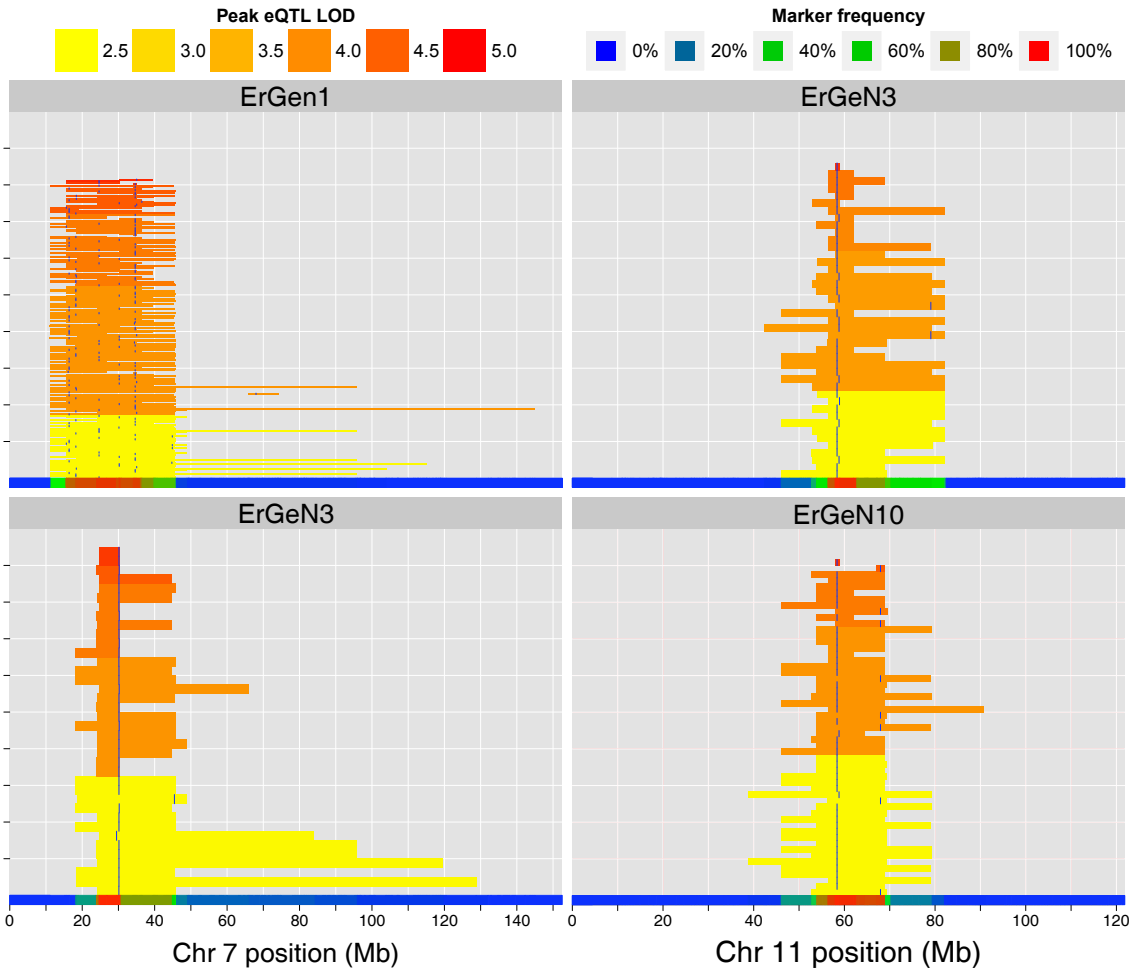

Figure 3.9. Support intervals for the major eQTL hotspot on Chr 7 for ErGeN1 and ErGeN3, and the eQTL hotspot on Chr 11 for ErGeN3 and ErGeN10. Each horizontal line represents an individual probe-set's 1.5 LOD-drop support interval, ordered and colored based on peak LOD score. Blue ticks indicate peak eQTL locations. The heatmap along the x-axis represents the percentage of probe-set support intervals that encompass the underlying markers. *Trans*-bands support intervals were defined as the chromosomal regions harboring at least 80% of the individual probe-set's eQTL support intervals. Full results are provided in Table 3.2.

The regulatory hotspot underlying the Chr 15 *trans*-band has previously been implicated as a regulator of two ethanol behavioral phenotypes, including an ethanol preference QTL mapped using congenic lines derived from B6 and BALB/cJ mice (Vadasz et al., 2000); as well as a QTL underlying loss of righting reflex (LORR) due to ethanol (Bennett et al., 2002; Markel et al., 1996). The primary candidate regulator of this *trans*-band was NEL-like 2 (*Nell2*) (Protein kinase C binding protein), which showed the highest regional response to ethanol. *Nell2* was an important hub of ErGeN1, as the network's fifth most central gene. While *Nell2*'s baseline transcription was strongly regulated by a *cis* eQTL, its ethanol response was modulated by the Chr 13 regulatory hotspot.

Similar to the Chr 7 *trans*-band, the regulatory hotspot on Chr 11 was linked to *trans*-bands from multiple networks, ErGeN3 and ErGeN10 (Table 3.2). Two strong candidate genes emerged from this region: *Gria1* and *Ncor1* (Table 3.3). From a hypothetical functional perspective, both genes are highly intriguing candidates; *Gria1*, as an ionotropic glutamate receptor and *Ncor1* as a transcriptional repressor acting through nuclear receptors and histone deacetylation. In our expression data, both genes were significantly ethanol-responsive, however, *Ncor1*'s response was stronger than *Gria1*'s. Furthermore, while the baseline expression of *Gria1* was primarily regulated by a highly significant *cis* eQTL, regulation of *Ncor1* was modulated by a suggestive *cis* and *trans* eQTL, the latter of which coincided with the Chr 7 *trans*-band. Reanalysis of *Ncor1*'s expression using a two-locus model revealed a significant interaction between the Chr 11 and Chr 7 eQTL (data not shown).

Table 3.4. Functional analysis of major ErGeNs

Functional category	Source	FDR p-value	# of genes
ErGeN1			
GTPase activity	GO:MF	1.5E-07	26/219
Regulation of synaptic transmission	GO:BP	1.85E-07	21/153
Neurotransmitter secretion	GO:BP	2.31E-06	14/85
Synapse part	GO:CC	3.08E-09	32/270
Dendrite	GO:CC	8.56E-09	25/182
Synaptosome	GO:CC	7.85E-07	15/91
PTEN pathway	MigDB	2.86E-06	7/18
ErGeN3			
RING-type zinc fingers	HGNC	1.2E-07	16/209
Synapse part	GO:BP	1.83E-06	16/270
FHF complex	GO:CC	2.84E-05	3/5
Histone deacetylase complex	GO:CC	3.56E-04	5/43
Potassium channels	HGNC	2.28E-05	6/88

3.6 Biological relevance of ethanol-responsive networks

3.6.1 Functional analysis of ethanol responsive networks

As was done for total ethanol-responsive gene sets, we investigated GO or pathway functional over-representation for the S-score networks. The vast majority of networks were over-represented for at least one gene family, protein domain/interaction, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway or GO category, significant at a FDR level of 5% (Table S8). ErGeN1 was strikingly enriched for proteins with GTPase activity (p-value = 1.5E-07), including *Rab3a*, which mediates ataxic consequences of ethanol consumption and influences ethanol preference (Kapfhamer et al., 2008). Both ErGeN1 and ErGeN3 were significantly enriched for genes encoding proteins that localize to the synapse (Table 3.4). In contrast, S-score networks 2 and 12 had a large over-representation of genes related to ribosome function and oxidative phosphorylation.

3.6.2 Phenotype correlation analysis

Using the BXD panel of mouse strains also allowed for direct comparison of ethanol gene expression data with the wide variety of phenotypic traits previously profiled in the BXD strains. To detect high-level phenotypes regulated by ethanol-responsive gene networks, we tested associations between ErGeNs and over 2,000 phenotypes available from GeneNetwork. This analysis was conducted by measuring correlations between GeneNetwork phenotypes and synthetic traits generated by principal component analysis (PCA) of ErGeN *trans*-bands. The first principal component of each *trans*-band was used for computational ease and clarity. Performing this analysis at the network and *trans*-band level made it possible to detect patterns of phenotypic associations with improved specificity. As expected, the analysis showed a striking clustering of *trans*-band

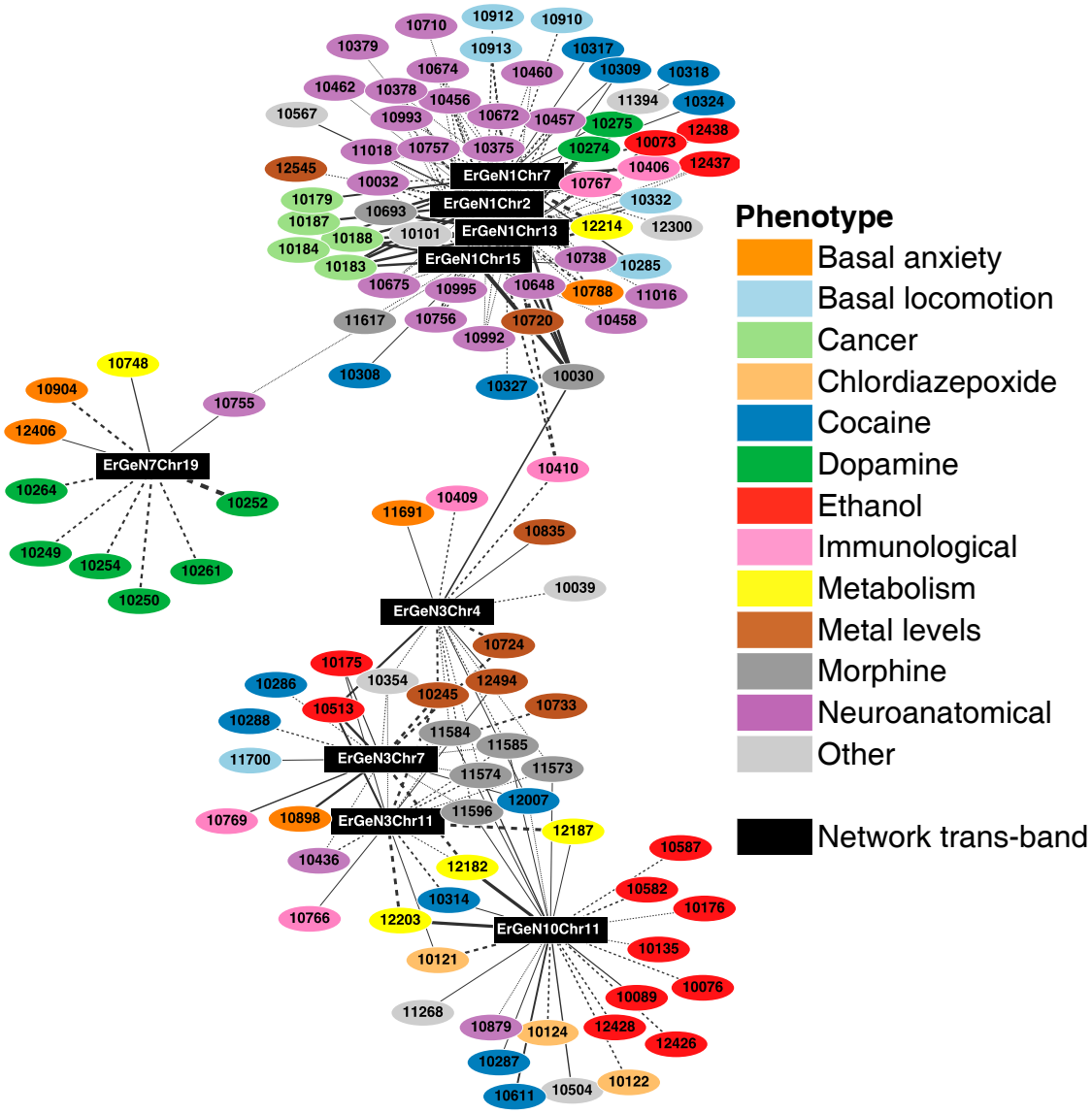

Figure 3.10. Correlations between ethanol responsive networks and phenotypes ErGeN *trans*-bands have distinct phenotypic correlations. Correlations between principal component traits of ErGeN *trans*-bands and BXD phenotypes (p-value < 0.01). Edge thickness indicates strength of network/phenotype association and dashed lines indicate a negative correlation. Phenotype nodes are labeled with trait IDs that can be queried on GeneNetwork.

for individual ErGeNs and associated phenotypes (Figure 3.10).

The GeneNetwork phenotype database contains a large number of neuroanatomical morphometric measurements (Martin et al., 2006). Many of these were strongly associated with ErGeN1 in its entirety (i.e., all *trans*-bands), including ventral hippocampus volume, overall brain weight, dorsal thalamus volume and amygdala basolateral complex volume. This network was also highly correlated with the β-max for naloxone binding (Belknap et al., 1995), a μ-opioid receptor antagonist that is an approved treatment for alcoholism. Whereas only a subset of ErGeN's *trans*-band were correlated with morphine metabolism rate (Wahlström et al., 1986); the same two *trans*-bands also correlated with ethanol acceptance in a two-bottle choice test (Crabbe et al., 1983).

This analysis also revealed ErGeN3 to be important potential mediators of phenotypic responses to several drugs of abuse. As a whole, ErGeN1 impacts both baseline locomotor activity (Yang et al., 2008) and habituation (Jones et al., 1999) in novel open field tests, but the effect of cocaine on these phenotypes was primarily correlated with ErGeN1's Chr 7 *trans*-band. Interestingly, non-locomotor based responses to cocaine were associated exclusively with ErGeN3, including measurements of stereotypic repeated movements (Jones et al., 1999; Tolliver et al., 1994) and conditioned place preference for the drug (Phillips et al., unpublished). Given the importance of dopamine levels in activating these behaviors, particularly stereotypy, we expected to find a strong connection between ErGeN1's Chr 7 *trans*-band and the dopamine binding phenotypes included in the GeneNetwork database. Instead, we observed that ErGeN7's solitary *trans*-band on Chr 19 to be the primary correlate of these measurements, which included *Drd1* and *Drd2* binding density in the dorsal striatum and NAc (Jones et al., 1999).

Along with ErGeN1, ErGeN3 was related to Naloxone β-max concentration but also showed strong correlations with morphine induced locomotor activation and naloxone induced morphine withdrawal (Phillips et al. unpublished). These morphine phenotypes

were also connected to ErGeN10. This overlap is perhaps not surprising given the strong association between many genes within ErGeN3 and ErGeN10 (Figure 3.3), as well as the shared *trans*-band support interval on Chr 11. However, one distinction between ErGeN3 and ErGeN10 was the clustered connections of numerous ethanol relevant phenotypes to ErGeN10. While ErGeN3 correlated with ethanol metabolism rate (Grisel et al., 2002) and blood glucose levels following ethanol treatment (Risinger, 2003), ErGeN10 appears more related to ethanol behavioral phenotypes, including ethanol induced locomotor activation (Crabbe et al., 1983), anxiolysis (Cook et al., unpublished) and sensitization (Cunningham, 1995).

3.7 Discussion

Here we have presented results from the first genetic analysis of gene networks that constitute transcriptional response of PFC to acute ethanol. Our analysis identified unique gene networks with implications on ethanol-evoked neuroadaptive mechanisms and behaviors, and showed that the response of such networks is governed by overlapping sets of discreet genetic loci. Perhaps most importantly, this analysis highlighted a series of hub genes as potentially major factors influencing brain responses to ethanol, setting the stage for future mechanistic studies and possible development of novel therapeutic approaches to alcoholism.

We leveraged the genetic variance in ethanol expression profiles by deriving dense paraclique gene networks co-regulated by acute ethanol. These networks likely represent initial perturbations of key molecular pathways, which, upon repeated consumption of ethanol, produce downstream neuroadaptations associated with alcohol abuse and dependence. The functional results of ErGeN1 and ErGeN3 (both of which were highly populated with robust ethanol-responsive genes) support this assertion, as both networks

were significantly enriched for proteins involved in neurotransmission and synaptic plasticity (Table 3.4, section B.8).

The major hub genes of PFC saline versus ethanol S-score networks, and particularly ErGeN3, included a number of genes previously implicated in drug dependence and neurological disease. The aforementioned node with the highest betweenness centrality in ErGeN3 was a probe-set targeting *Grm3*. It is well established that metabotropic glutamate receptors play an important functional role in the development of AUD (Chandler et al., 2006; Gass and Olive, 2008; Vengeliene et al., 2008). Studies have demonstrated, in particular, that modulation of *Grm3* decreases ethanol seeking in rats (Bäckström and Hyytiä, 2005); although the agonists used in these studies also bind *Grm2*. *Grm3* is also a high priority candidate gene for schizophrenia, as a group II mGluR agonist (LY354740) blocked many symptoms induced in the rat phencyclidine treatment model of schizophrenia (Moghaddam and Adams, 1998). *Grm3* has also been associated with schizophrenia phenotypes in human GWA studies (Egan et al., 2004). Among the genes adjacent to *Grm3* in ErGeN3, the strongest correlation was between *Grm3* and *Nrg3* ($r = 0.97$, p-value $<$ 1e-16). Like *Grm3*, *Nrg3* is a highly connected gene in ErGeN3 as well as a schizophrenia candidate gene (Kao et al., 2010; Morar et al., 2011).

The large conductance potassium channel, *Kcnma1*, is also an ErGeN3 hub gene (Figure 3.7). In addition to its known functional response to ethanol exposure (Dopico et al., 1996), *Kcnma1* is a very intriguing hub gene because it is a proven regulator of acute ethanol induced intoxication in *C. elegans* (Davies et al., 2003). Furthermore, two recently published human GWA studies have provided preliminary evidence for a link between *Kcnma1* and alcohol dependence (Edenberg et al., 2010; Kendler et al., 2011). The study by Kendler et al. also identified another voltage gated potassium channel, *Kcnq5*, as having an association with alcohol dependence. This is an exciting

result, as *Kcnq5* is directly adjacent to *Kcnma1* in ErGeN3, and both genes are highly ethanol-responsive and major hubs of the network.

In addition to identifying hub genes as leading candidates for future verification studies, our genetic dissection of ethanol-responsive gene networks also produced clues regarding the mechanisms underlying ethanol network responses. Identification of chromosomal hot spots linked to ethanol responses for entire gene networks provides genetic evidence for hubs influencing the response of ErGeN's and expands our understanding of brain molecular signaling events responding to ethanol. For example, the sodium channel *Scn1b* was a hub gene in ErGeN1, showed robust ethanol-responsiveness, had a highly significant *cis* eQTL and also was a strong candidate for regulating a trans-band of ErGeN3 mapping to exactly the location of *Scn1b*. *Scn1b* codes for a regulatory subunit of sodium channels which are crucial to action potential propagation. Ethanol has been shown previously to inhibit sodium channel function (Horishita and Harris, 2008). This data suggests that *Scn1b* and other such potential regulators of ethanol-responsive trans-bands may be key modulators for extensive portions of the overall ethanol responsome.

Chapter 4

Anxiolytic-like response to acute ethanol

4.1 Behavioral responses to acute ethanol

The discovery that responses to acute ethanol are powerful predictors of an individual's risk for alcoholism was a major breakthrough in alcohol genetics research. In a landmark study, Schuckit (1994) found that college students less affected by the intoxicating effects of acute alcohol, so called low-responders, were significantly more likely to develop alcohol dependence later in life. Furthermore, students with a family history of alcoholism qualified as low-responders far more frequently than students with no close relatives suffering from alcoholism; a clear indication that acute ethanol responses are largely impacted by genetic factors. This inverse relationship between level of response to acute ethanol and long-term drinking behaviors has been replicated in mouse models, making it possible to investigate the underlying mechanisms controlling acute ethanol sensitivity. For example, given a two-bottle choice, B6 mice will consume more ethanol than D2 inbred mice, which are largely uninterested in the voluntarily consumption of ethanol but are much more sensitive to ethanol induced acute locomotor activation (Metten et al., 1998). Numerous, but not all, studies in various gene targeting models

also show that drinking behavior tends to vary inversely with such acute responses such as locomotor activation or sedation.

4.1.1 Anxiety as a risk factor for alcoholism

Anxiety has long been considered a risk factor for alcoholism. One early theory, referred to as the tension reduction hypothesis, postulated that ethanol's anxiolytic effect would be more rewarding for anxiety prone individuals and would reinforce the benefits of drinking, ultimately leading to more frequent consumption (Cappell and Herman, 1972). While evidence supporting this relationship in humans is lacking, the considerable comorbidity between alcohol use disorders and anxiety disorders (Kessler et al., 1994), as well as the observation that alcoholics frequently report anxiety reduction motivates them to drink (Newlin and Thomson, 1990), strongly suggests that anxiety is indeed an important contributor to an individual's risk for developing alcoholism.

The tension reduction hypothesis has found support in studies of rodent anxiety models, which have demonstrated that highly anxious rats exhibit increased ethanol preference and consumption (Spanagel et al., 1995) and both rats and mice treated with ethanol will spend significantly more time in the open arm of an elevated plus-maze (Boehm et al., 2002; LaBuda and Fuchs, 2000, 2001), which is interpreted as a reduction in anxiety. Despite the apparent link between anxiety and alcoholism, and the numerous studies demonstrating ethanol induced anxiolysis can be consistently measured with rodent models, there is a dearth of studies investigating the genetic components underlying the anxiolytic response to acute ethanol.

4.1.2 Initial mapping of *Etanq1*

Dr. Putman focused his thesis project on investigating the mechanisms underlying the anxiolytic effects of acute ethanol using the light-dark transition model of anxiety (Crawley and Goodwin, 1980; Putman, 2008). As nocturnal creatures, rodents are naturally averse to bright lights. Therefore, in the light-dark model, the light side of the chamber acts as a stressor and anxiety is measured by analyzing the percent time spent in the light (PTS) or the percent distance traveled in the light (PDT). An additional benefit of this model is total locomotor activity (TLA) can also be measured, which is a phenotype that is also relevant to alcohol research (Boehm et al., 2002; Lessov et al., 2001).

His work revealed that both B6 and D2 mice traveled significantly further in the light side of the chamber following ethanol treatment than saline treated controls. Furthermore, the degree of anxiolysis exhibited by the B6 mice appeared larger than the D2 mice. Dr. Putman expanded the light-dark box studies across the BXD RI panel, making it possible to perform gene mapping studies by identifying chromosomal regions that influenced ethanol response in a genotype specific manner. This ultimately led to the identification of ethanol-induced anxiolytic-like response QTL 1 (*Etanq1*), a strong mediator of ethanol's impact on anxiety-like phenotypes.

4.2 Extending the *Etanq1* project

This project was concerned with extending his work by identifying the causative QTG underlying *Etanq1*, which would provide novel insight into the molecular mechanisms of ethanol-induced anxiolysis. While the ratio of mapped behavioral QTL to successfully identified QTG has been uninspiring, I believed the project had a higher probability of

success for several reasons. First, the effect size of *Etanq1* is quite large, with the QTL accounting for over 30% of the variance in ethanol-induced anxiolysis as measured by PDT. A 2005 review of rodent QTL mapping studies conducted a survey of behavioral QTLs and found the average effect size for uncloned QTL is 5.8% while the average effect for the small number of successfully cloned QTL is 26% (Flint et al., 2005), which highlights the important relationship between QTL effect size and likelihood of success. Second, the PFC, NAc and VMB microarray expression datasets described in Chapter 2 were derived from the same mice used in the light-dark box (LD box) behavioral assay. Through the integration of these data it was possible to perform genetic correlations with transcript variation and prioritize genes within the *Etanq1* support interval associated with ethanol-induced anxiolysis.

4.3 Preliminary studies

4.3.1 Validity of light-dark box model of anxiety

The light-dark transition model of anxiety was selected to investigate the anxiolytic-like response to acute ethanol. The construct validity of this model was confirmed by comparing the results of D2 mice that received IP injections of either 0.9% saline or diazepam (2mg/kg), a benzodiazepine that produces anxiolysis and central nervous system (CNS) depression through activation of $GABA_A$ receptors (Sieghart, 1994). Mice in the diazepam treatment group traveled significantly further (p-value = 0.0102) and spent significantly more time (p-value = 0.0176) in the light side of the light-dark box than the saline controls. Any increase in percent time or percent locomotor activity in the light is interpreted as a reduction in anxiety.

4.3.2 Anxiolytic-like response to acute ethanol in B6 and D2

The LD box method provides a valid measurement of pharmacologically induced anxiolysis in mice, as demonstrated by the diazepam experiment above. While other measures have been used to measure anxiety-like behaviors in rodents, Drs. Putman and Miles chose the LD box due to the relatively high-throughput nature of the assay and its reproducibility. Figure 4.1 provides LD box data from initial experiments investigating the anxiolytic-like effects of acute ethanol in B6 as well as D2 mice. Mice received an IP injection of 0.9% saline or a sedative dose of ethanol (1.8 g/kg). As with the diazepam experiment, mice treated with ethanol spent significantly more time in the light and traveled significantly further in the light. While both strains demonstrated a significant anxiolytic response to ethanol, the magnitude of the B6 response was stronger across both anxiety measures. In contrast, only the D2 mice significantly increased their TLA following ethanol treatment (Figure 4.1, right panel), which is consistent with previous reports (Phillips et al., 1995) and suggests that PDT and PTS are influenced by different genetic mechanisms than the drivers of ethanol-induced locomotor activation.

4.3.3 Provisional QTL for ethanol anxiolytic-like response

To identify genetic influences underlying the anxiolytic-like response to ethanol, the LD box assay was repeated across 27 strains from the BXD RI panel following saline or ethanol IP injections. Of the 27 assayed strains, 23 exhibited a robust anxiolytic-like response to acute ethanol (Figure 4.2). As expected, variation in the measured behaviors followed a continuous distribution, indicating that these phenotypes are complex traits influenced by multiple polymorphic loci. Interval mapping was carried out for PDT, PTS and TLA following saline and ethanol treatment using the same set of genotypic data described in section 3.3.1. These results are summarized in Table 4.1. The anxiolytic-like

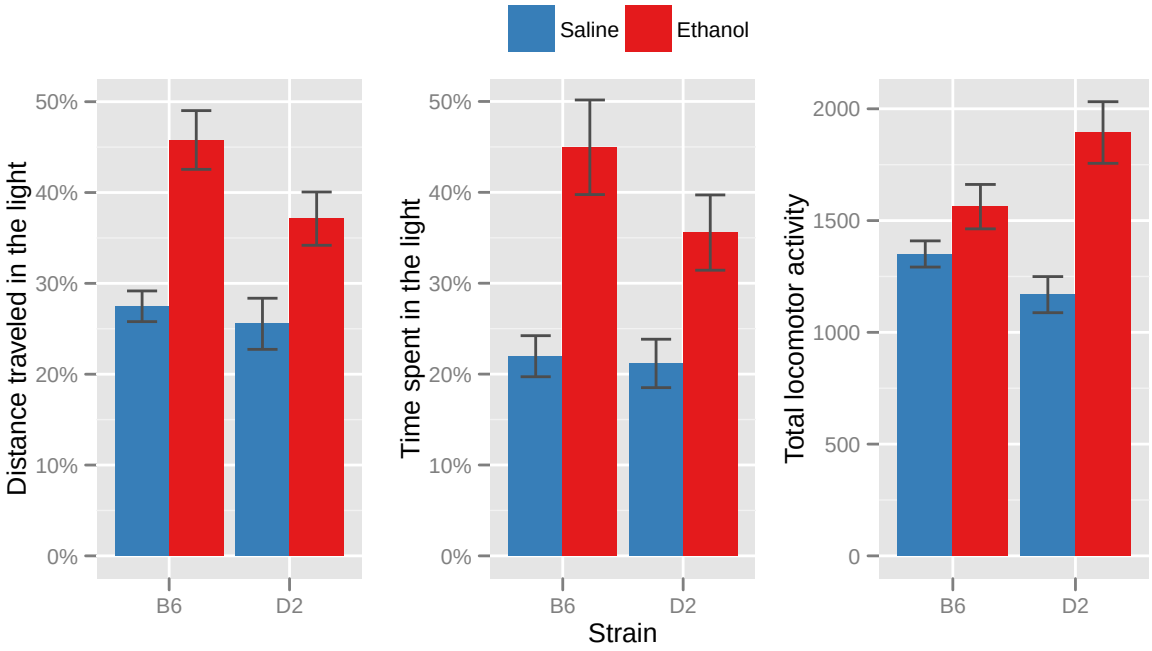

Figure 4.1. B6 and D2 mice anxiolytic-like behaviors. Percent distance traveled in the light (left) and total locomotor activity (right) following IP injection of saline (blue) or ethanol (red) across BXD progenitors, B6 and D2 (n = 13–16). Bars represent mean ± SEM. A two-way ANOVA followed by Tukey's HSD post-hoc analysis showed that ethanol significantly increased the PDT for B6 (p-value < 0.001) and D2 (p-value < 0.05). However, only D2 mice exhibited a significant increase in TLA following ethanol.

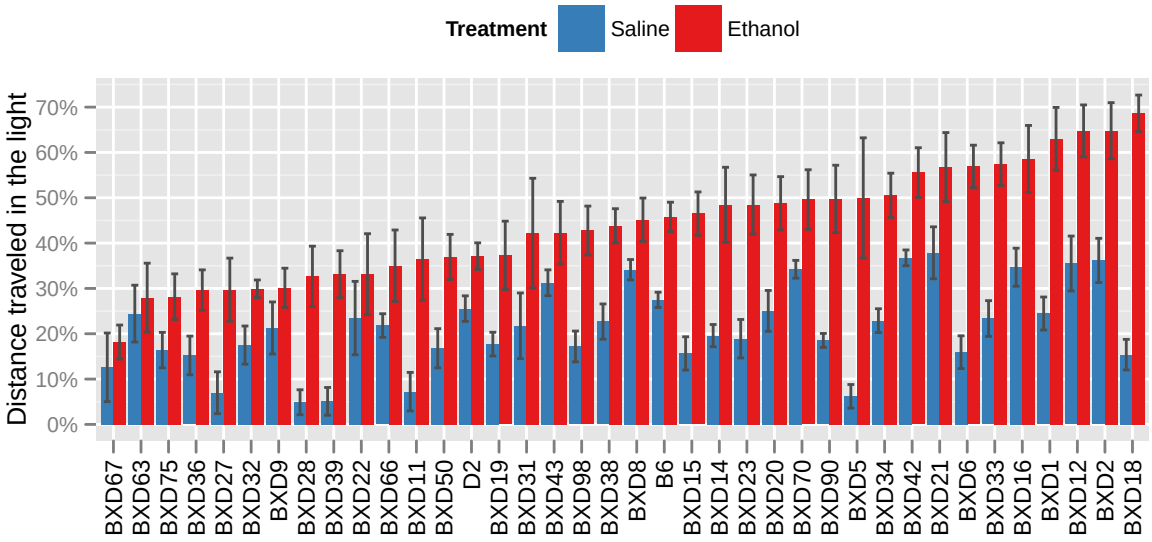

Figure 4.2. Variation in the anxiolytic-like response to acute ethanol across B6, D2 and BXD strains assayed as part of the provisional mapping, confirmation and fine-mapping of *Etanq1*.

behaviors, PDT and PTS, were highly correlated in both the saline ($r = 0.93$, p-value $= 3 \times 10^{-13}$) and ethanol ($r = 0.98$, p-value $= 2 \times 10^{-16}$) treatment groups and consistently produced overlapping QTL, indicating that these phenotypes are measuring the same underlying constructs. As PDT produced the more significant QTL, it will be the only anxiety metric discussed henceforth.

While the majority of strains exhibited a robust increase in locomotor activity post-ethanol, the primary genetic driver of locomotor variation was the same for both treatments as evidenced by the significant QTL for TLA on the same distal region of Chr 1 following saline or ethanol (Table 4.1). The anxiety measures also produced QTLs on Chr 1, however their peak locations relative to the locomotor QTL were approximately 20 Mb proximal following saline and 30 Mb distal following ethanol. Though support intervals for the anxiety behaviors and TLA overlap, further studies are necessary to determine whether these phenotypes are being influenced by a common locus. Interestingly, the peak location of the TLA QTL on Chr 1 coincides with the QTL hotspot we referenced in

Table 4.1. Provisional QTL mapped for anxiety and locomotor phenotypes

Phenotype	Chr	Marker	Peak (Mb)	Support interval (Mb)	LOD
Saline PDT	11	rs13481251	116.91	4.408–121.409	3.32
Ethanol PDT	12	rs3716547	68.67	53.911–71.6617	5.16
Saline TLA	11	rs13481117	79.07	4.408–121.409	3.11
Saline TLA	9	mCV23098764	51.52	10.573–124.039	3.37

section 1.5.2, which modulates a large number of phenotypes, including several ethanol phenotypes, and was recently dissected by Mozhui et al. (2008).

The most striking result from the interval mapping analysis is the linkage of anxiety behaviors following ethanol to Chr 12 (Figure 4.3). The ethanol-induced anxiolytic-like response QTL 1 identified for PDT following ethanol treatment reaches a peak LOD score of 5.16. Using 10,000 permutations to derive an estimate of the genome-wide false positive rate revealed that *Etanq1* is highly significant (p-value < 0.01) according to the criteria established by the Complex Trait Consortium (Abiola et al., 2003). *Etanq1* accounts for 35.5% of the phenotypic variance, with the B6 allele increasing the ethanol-induced anxiolytic effect (Sen et al., 2007). We used Bayes credible intervals with 97% coverage to obtain *Etanq1*'s support interval location (Manichaikul et al., 2006), which extends from 53.73 Mb to 71.47 Mb across Chr 12 and contains 106 genes.

4.3.4 Confirmation of *Etanq1*

To confirm the linkage between *Etanq1* and ethanol-induced anxiolysis, the LD box assay was repeated with 6 additional BXD strains from the independently derived panel of AI lines (Peirce et al., 2004) and supplied from Oak Ridge National Laboratory, rather than Jackson Laboratory, which supplied the 27 strains used in the provisional genetic screen. These strains were split into samples based on treatment and genotype at *Etanq1*'s peak marker, rs13481514. As shown in Figure 4.4, only the group carrying a B6 allele

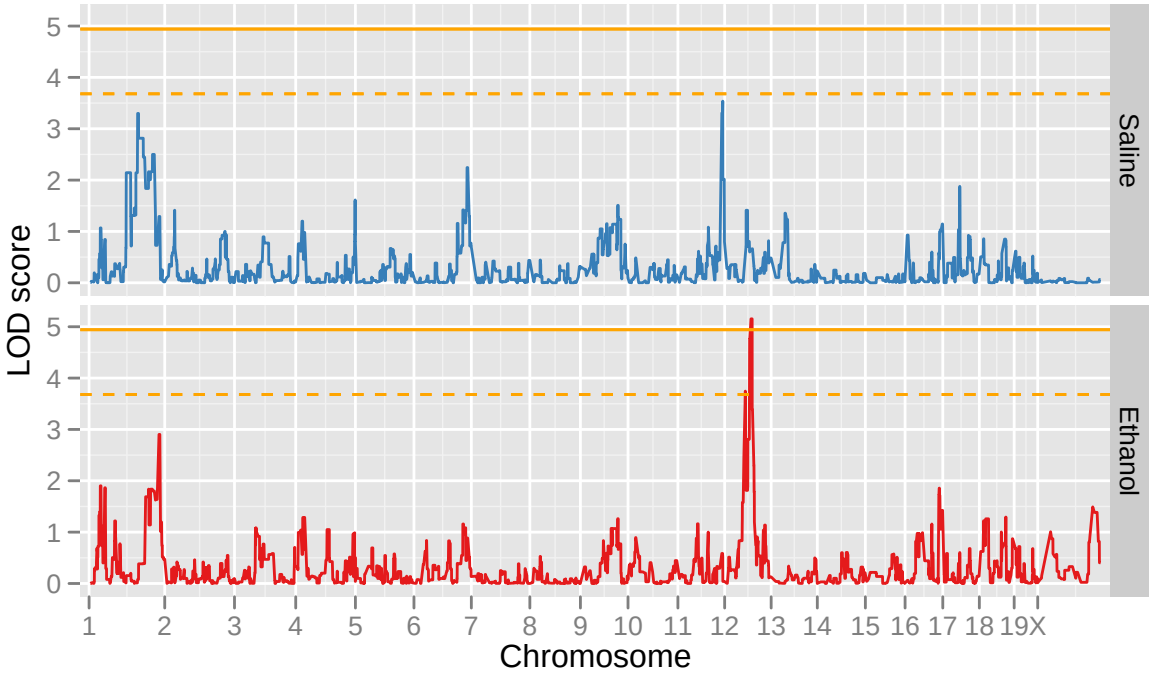

Figure 4.3. Provisional QTL for anxiolytic-like response to acute ethanol Genome-wide interval mapping results for PDT across 27 BXD RI strains following treatment with saline (blue) or ethanol (red).

at *Etanq1* exhibited a significant increase in PDT following ethanol treatment ($p = 5 \times 10^{-6}$). The D2 *Etanq1* strains did travel slightly further in the light after ethanol but the difference was non-significant compared to the saline treated group ($p = 0.37$).

4.4 Methods

4.4.1 Mice

Details about mice used in these studies are provided in section 2.3.1 on page 35.

4.4.2 Light-dark box behavioral assay

Behavioral testing was conducted as described by Putman (2008). Briefly, all animals were tested between 10:00 A.M. and 1:00 P.M. Following a 1 hour acclimation period to the behavioral testing room, animals were restrained in a 50 ml conical tube for 15 minutes followed by IP injections with either physiological saline (0.9%) or 1.8 g/kg ethanol (12.8% w/v) in 0.9% saline. The restraint stress was employed to create an artificial baseline level of anxiety-like behavior. This method controlled for environmental perturbations of anxiety-related behavior in individual mice such as social stress.

Following a 5 minute delay from the time of injection, each animal was placed in the center of the light chamber facing the entrance to the dark chamber of the LD box. Once the animal entered the dark compartment, anxiety-like scores were collected in 5 minute intervals for a total of 10 minutes. Behavioral measures were recorded in both chambers of the light-dark box and included distance traveled, time spent and TLA. Anxiety-like measures were reported as PTS and PDT to control for locomotor activity. An increase in either measure was interpreted as a reduction in anxiety-like behavior.

The LD box consists of two equally sized compartments (30 cm × 30 cm × 15

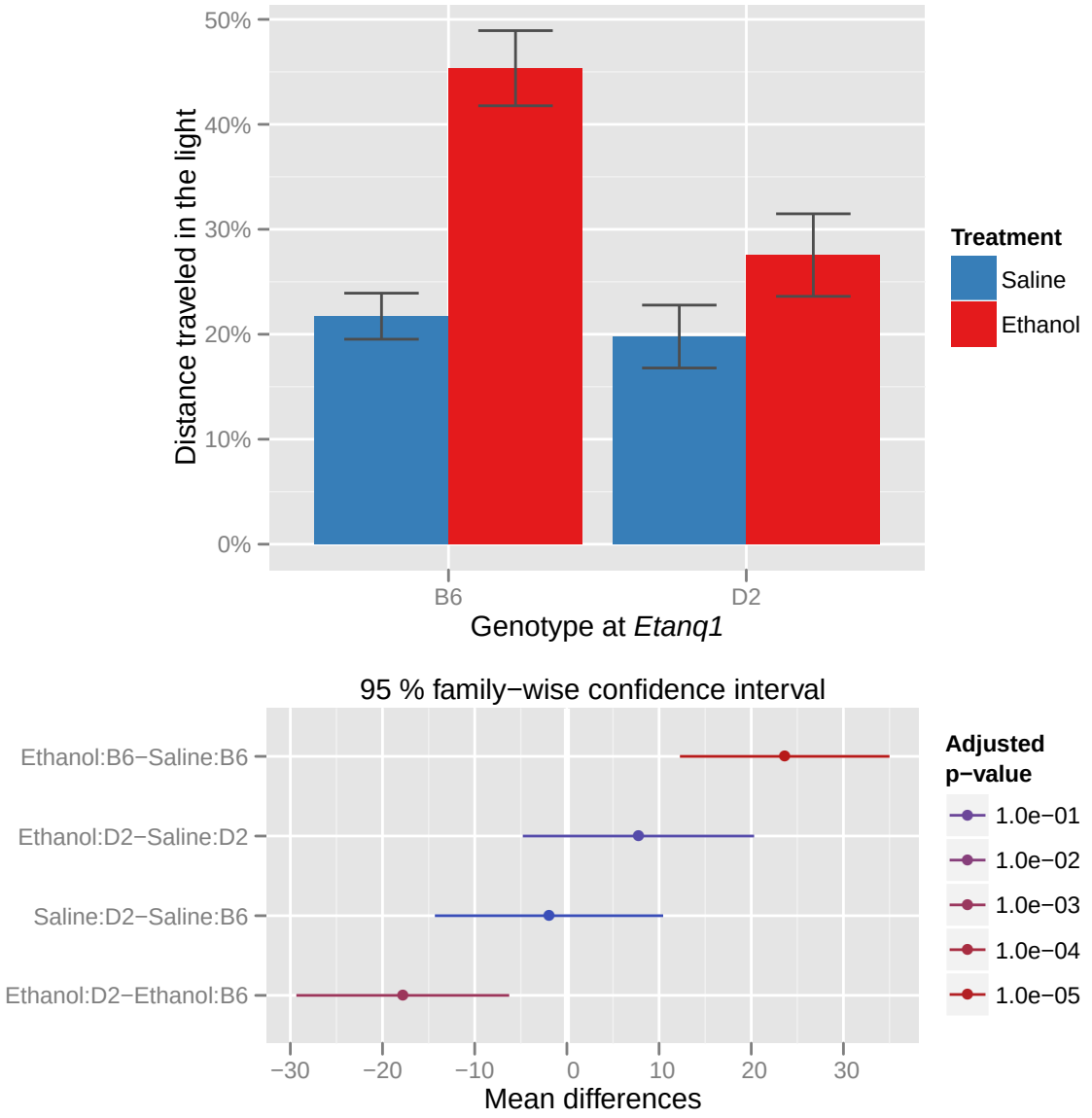

Figure 4.4. Confirmation of *Etanq1*. PDT following saline (blue) and ethanol (red) for 6 novel BXD strains (n = 4–7), collapsed into groups based on genotype at *Etanq1*'s peak marker location on Chr 12. An ANOVA with Tukey's HSD *post-hoc* analysis revealed a significant allelic effect on this behavior following ethanol treatment (p-value < 0.0001), with the B6 allele causing a significant increase in the anxiolytic response.

cm) separated by a black plastic partition with an opening in the middle to allow for light-dark transitions (Med Associates Inc., St. Albans, VT, USA). The box was enclosed in a sound-attenuating box equipped with overhead lighting and fan ventilation. The system was interfaced with Med Associates software enabling automatic measurement of activity using a set of 16 infrared beam sensors along the X-Y plane.

4.5 Fine-mapping *Etanq1*

4.5.1 Intra-researcher reliability

Measurements of mouse behavioral phenotypes are extremely sensitive to environmental effects. Even when great care is taken to standardize testing procedures and environmental conditions across different laboratories, identical strains of mice put through an identical battery of assays can yield significant laboratory-specific differences (Crabbe et al., 1999). The researcher handling the mice could be an important factor driving these differences. This potential source of error was particularly relevant to my thesis, since fine-mapping *Etanq1* required additional BXD strains by a different researcher. Therefore, to ensure the behavioral measurements I generated for the additional BXD strains could be integrated with the original data obtained by Dr. Putman, I repeated the LD box assay with B6 mice and compared our respective results, which are presented in Figure 4.5.

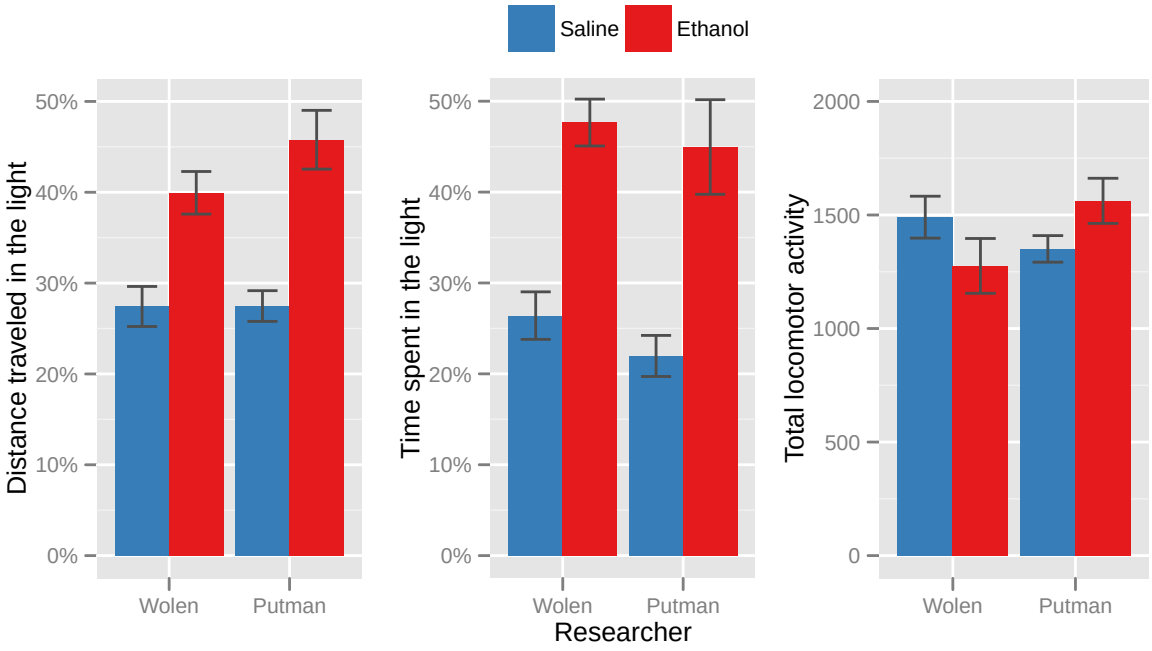

Figure 4.5. Reproducibility of ethanol-induced anxiolysis across researchers. LD box measurements of PDT (left), PTS (middle) and TLA (right) for B6 mice obtained by the author and Dr. Alex Putman. Red and blue bars represent the means of saline and ethanol treated mice, respectively. Error bars indicate the mean ± SEM. For both researchers a significant increase in PDT and PTS was observed following ethanol treatment. Significance was determined by ANOVA followed by Tukey's HSD *post-hoc* test, the results from which are provided in Tables 4.2, 4.3 and 4.4.

While some variation between researchers is to be expected, true intra-researcher reliability requires recreating the significant effect of acute ethanol treatment on the LD box anxiety measures without introducing significant researcher-specific bias. Indeed, performing a 2-way ANOVA with treatment and researcher as factors revealed no significant researcher effect was for PDT ($F_{(1,43)} = 1.17, p = 0.28$), PTS ($F_{(1,43)} = 0.73, p = 0.39$) or TLA ($F_{(1,43)} = 0.64, p = 0.43$). A Tukey's HSD *post-hoc* test was also performed to observe researcher-specific treatment effects across the three measures. Results for relevant PDT, PTS and TLA comparisons are provided in Tables 4.2, 4.3 and 4.4, respectively.

As expected, ethanol treated mice spent significantly more time (p-value = 1.46e-02) and traveled significantly further (p-value = 3.63e-02) than their saline treated counter parts. Thus, I was able to reproduce the anxiolytic-like response to acute ethanol in B6 mice. While the magnitude of these effects were slightly smaller than Dr. Putman's, this may be partially explained by differences in sample size; as my treatment groups contained only 8 individuals, whereas his contained 15–16. Still, our 95% CIs for the true difference between saline and ethanol treatment groups largely overlapped (Figure 4.6) and importantly, no significant effects were observed between researchers within the saline or ethanol treatment comparisons. That is, there were no statistical differences between our measures of PDT (saline p-value = 1.0, ethanol p-value = 0.43), PTS (saline p-value = 0.87, ethanol p-value = 0.96) or TLA (saline p-value = 0.72, ethanol p-value = 0.15). As such, the results of this analysis indicated I could contribute to pool of BXD LD box data generated by Dr. Putman, reasonably confident that the behavioral data generated for these novel strains would be statistically identical, regardless of which one of us performed the assay. And it was critical that this was established before we proceeded with the fine-mapping of *Etanq1*.

Table 4.2. Intra-researcher reliability: percent distance traveled in the light

Comparison		Mean difference	Lower CI	Upper CI	Adjusted p-value
Saline:Putman	Saline:Wolen	0.05	-10.38	10.49	1.00e+00
Ethanol:Putman	Ethanol:Wolen	5.84	-4.48	16.16	4.39e-01
Ethanol:Wolen	Saline:Wolen	12.52	0.60	24.43	3.62e-02
Ethanol:Putman	Saline:Putman	18.31	9.74	26.87	5.61e-06

Post-hoc comparisons made using Tukey's HSD for PDT. Similar results are provided for PTS and TLA in Tables 4.3 and 4.4, respectively.

Table 4.3. Intra-researcher reliability: percent time spent in the light

Comparison		Mean difference	Lower CI	Upper CI	Adjusted p-value
Saline:Putman	Saline:Wolen	-4.44	-20.15	11.27	8.74e-01
Ethanol:Putman	Ethanol:Wolen	-2.69	-18.23	12.85	9.67e-01
Ethanol:Wolen	Saline:Wolen	21.24	3.30	39.18	1.46e-02
Ethanol:Putman	Saline:Putman	22.99	10.09	35.89	1.25e-04

Table 4.4. Intra-researcher reliability: total locomotor activity

Comparison		Mean difference	Lower CI	Upper CI	Adjusted p-value
Saline:Putman	Saline:Wolen	-139.74	-497.63	218.14	7.25e-01
Ethanol:Putman	Ethanol:Wolen	286.45	-67.53	640.42	1.50e-01
Ethanol:Wolen	Saline:Wolen	-214.52	-623.26	194.21	5.05e-01
Ethanol:Putman	Saline:Putman	211.66	-82.13	505.46	2.33e-01

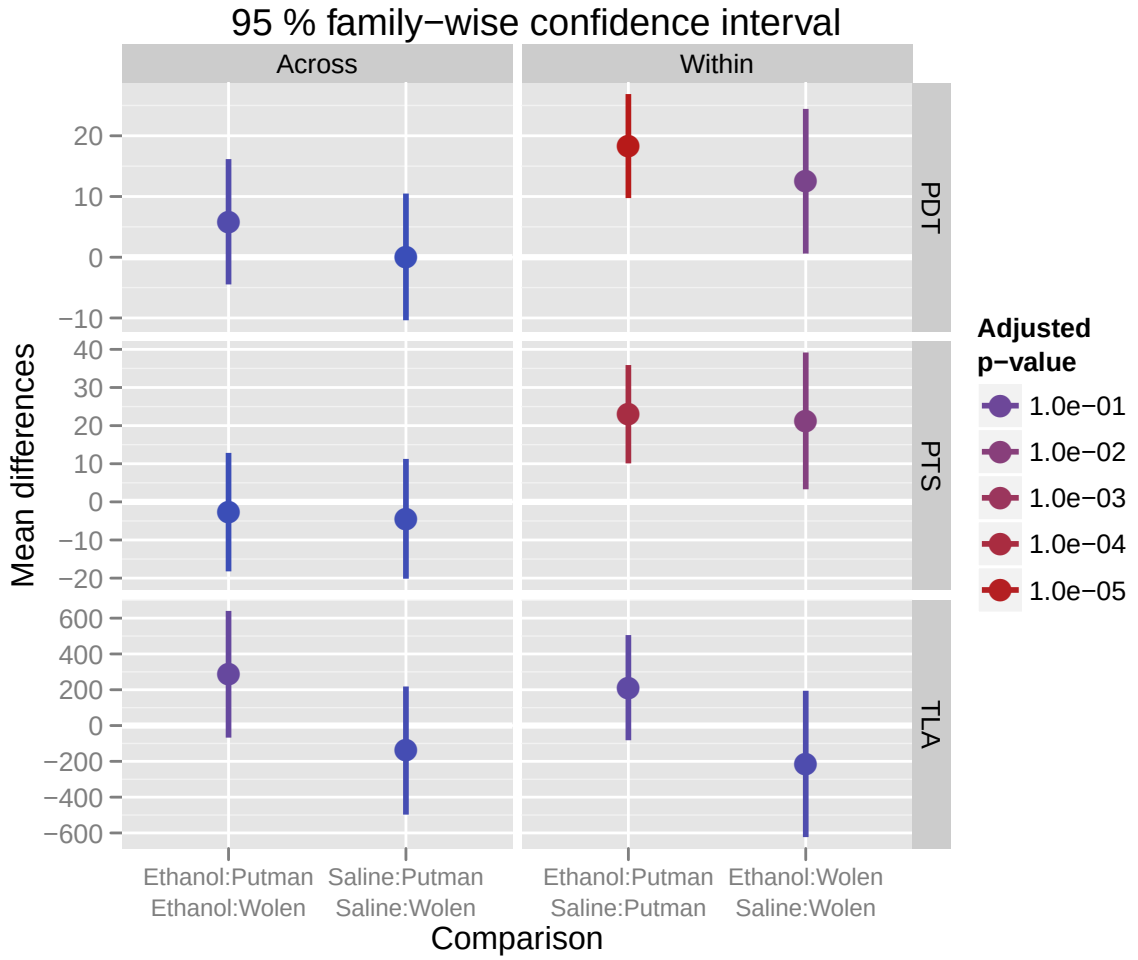

Figure 4.6. Tukey's HSD comparisons made for intra-researcher reliability ANOVA. Visualization of results from Tukey's HSD *post-hoc* analysis of the ANOVA performed to identify systematic differences in the LD box assays performed by the author and Dr. Alex Putman. Each point represents the mean difference between treatment groups with lines indicating the 95% CIs. Warmer colors correspond to lower adjusted p-values.

Table 4.5. Fine-mapped QTL mapped for anxiety and locomotor phenotypes

Phenotype	Chr	Marker	Peak (Mb)	Support interval (Mb)	LOD
Ethanol PDT	1	rs4222763	165.32	25.632–188.966	3.07
Ethanol PDT	12	rs13481514	70.70	69.125–72.561	5.99
Saline TLA	1	CEL-1_152747565	154.63	25.632–188.966	3.97
Ethanol TLA	2	rs6209325	147.88	5.767–181.542	3.27
Ethanol TLA	9	rs3656996	52.00	10.573–124.039	3.03

4.5.2 Assaying novel BXD strains

Our strategy for refining *Etanq1* involved assaying additional BXD strains that carry informative recombinations within the *Etanq1* support interval. Specifically, we sought to add novel BXD strains derived from the advanced intercross conducted by Peirce et al. (2004). Their breeding strategy entailed many additional rounds of intercrossing prior to inbreeding. As a result, these novel BXD strains contain twice as many reomcinbation events, providing greater genetic resolution for QTL mapping. An initial haplotype analysis of all novel BXD strains revealed that several strains carried a combination of B6 and D2 haplotypes in the region of interest. We predicted that including these strains in the QTL analysis would greatly enhance the genetic mapping resolution. Figure 4.7 displays the genotype distributions for all available BXD strains carrying at least one recombination event within the plotted region and indicates the novel strains that were chosen for fine-mapping.

4.5.3 Fine-mapped *Etanq1* QTL analysis

As predicted, including the BXD strains from the *Etanq1* confirmation experiment and the novel strains chosen for fine-mapping greatly enhanced the genetic mapping resolution. Repeating strain-mean interval mapping for PDT following acute ethanol treatment increased *Etanq1*'s peak LOD score to 5.99 and shifted its peak linkage

Figure 4.7. BXD genotypes across *Etanq1* support interval. Only BXD strains harboring recombination events within the region are included. Each strain's sequence of alleles runs horizontally along the grid. Strain order was determined through hierarchical clustering of allele sequences based on their Euclidean distance. Shapes along the left axis indicate whether a particular strain was assayed as part of the provisional mapping phase, the confirmation study or the fine-mapping effort. Black ticks along the x-axis indicate genetic marker location.

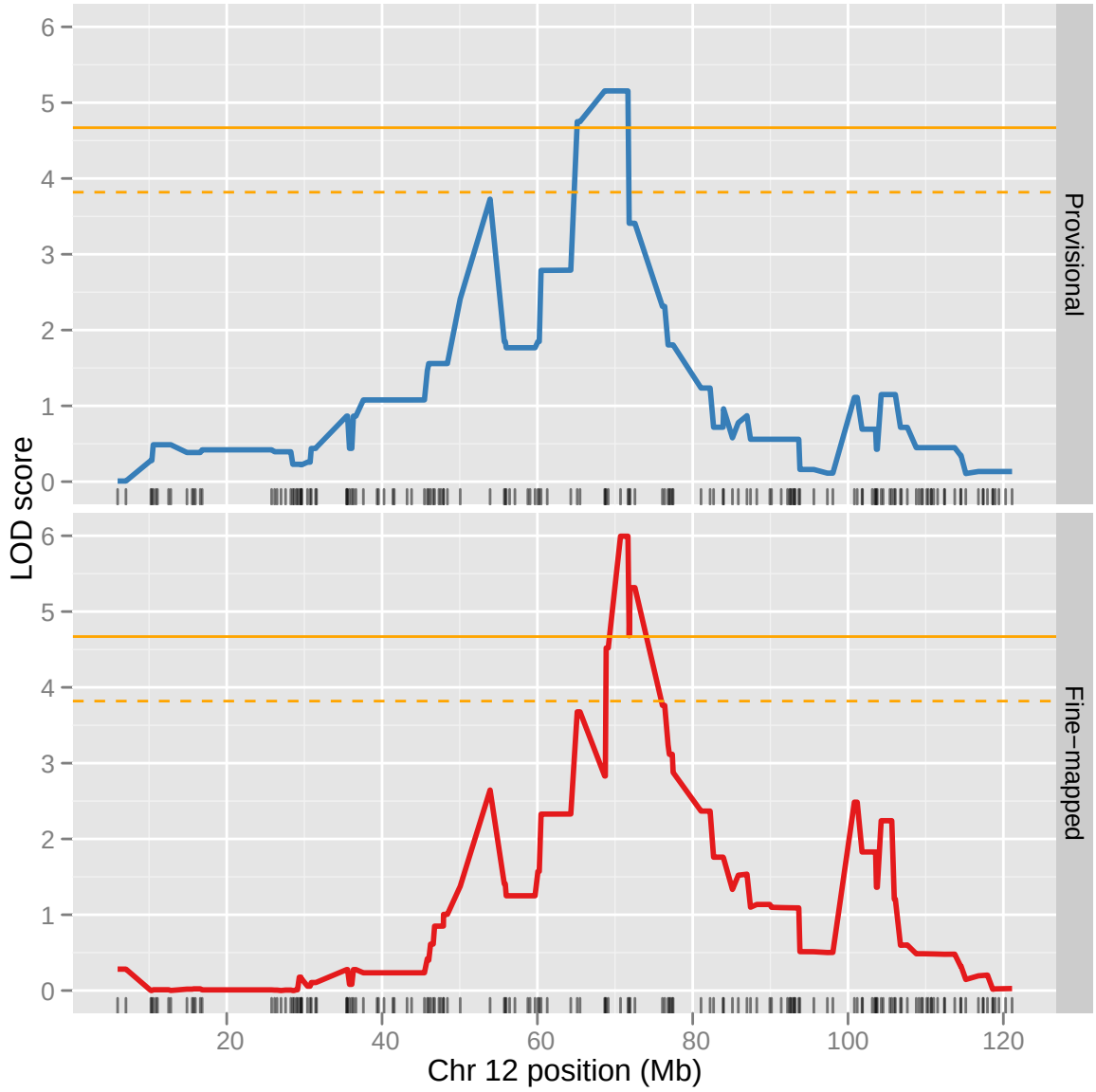

Figure 4.8. *Etanq1*'s provisonal and fine-mapped QTL across Chr 12. Strain mean interval mapping results for original 27 BXD strains (top) and expanded sample (red). The additional BXD strains caused the QTL's peak to narrow and proximally shift its region of peak association. Dashed and solid horizontal lines indicate genome-wide p-values of 0.05 and 0.01, respectively. The location of genetic markers used in this analysis are indicated by the vertical dashed plotted along the x-axis.

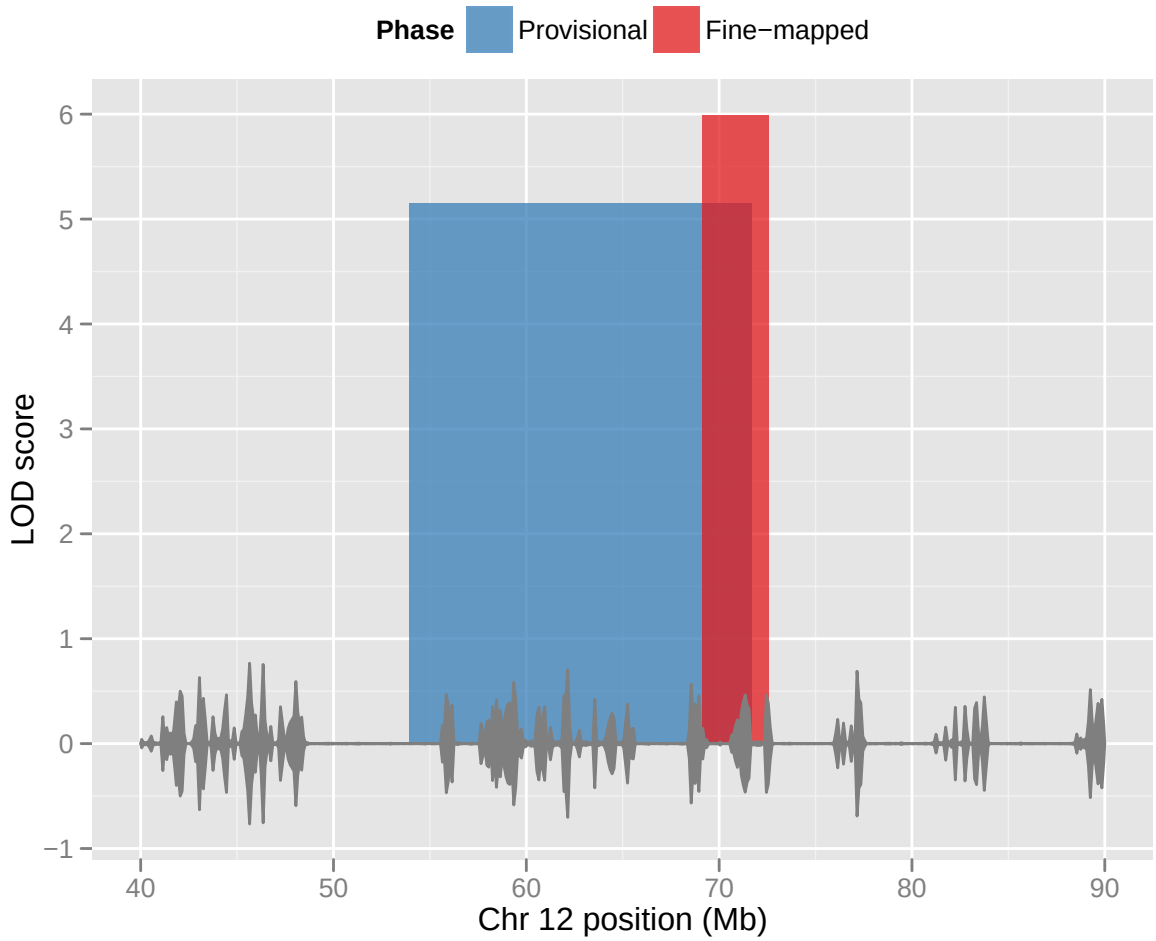

Figure 4.9. *Etanq1*'s fine-mapped support interval The original (blue) and fine-mapped (red) 97% support intervals for *Etanq1* across a subsection of Chr 12. Grey seismograph across the x-axis indicates the number of polymorphisms per kilobase (1×20^{-1}) between B6 and D2.

location proximally by ≈ 2 Mb (Figure 4.8). Most encouragingly, these additional strains substantially narrowed *Etanq1*'s support interval. While *Etanq1*'s provisional support interval spanned nearly 18 Mb, the fine-mapped support interval spans only 3.4 Mb (Table 4.5).

A comparison of the provisional and fine-mapped support intervals is provided in Figure 4.9. Using D2 genome sequence information, we counted all B6/D2 SNPs across Chr 12 and plotted the densities along the x-axis of Figure 4.9. Doing so revealed which regions are polymorphic and which are identical by descent (IBD) between the B6 and D2. More specifically, the fine-mapped *Etanq1* support interval appears to be centered around a highly polymorphic region flanked by areas that are nearly IBD. While the entire support interval should be explored by candidate QTGs, this highly polymorphic region between 70.9–71.5 Mb is of particular interest.

4.5.4 Haplotype analysis of refined *Etanq1* support interval

Following the success of the initial fine-mapping effort described above, we examined the haplotype structure of the *Etanq1* region. The purpose of this analysis was to determine whether continuing to assay additional BXD strains would provide substantive gains in genetic mapping resolution. The r^2 measure of LD was calculated for all pairwise combinations among the 162 genetic markers across Chr 12, using genotype data for all 93 BXD strains. The haplotype map produced from this analysis is presented in Figure 4.10. Superimposed on this map are the provisional and fine-mapped *Etanq1* support intervals.

The results from this analysis suggest how *Etanq1*'s support interval could be narrowed so dramatically with the addition of relatively few novel strains. If we roughly define a haplotype block as a chromosomal segment in which all genetic markers have

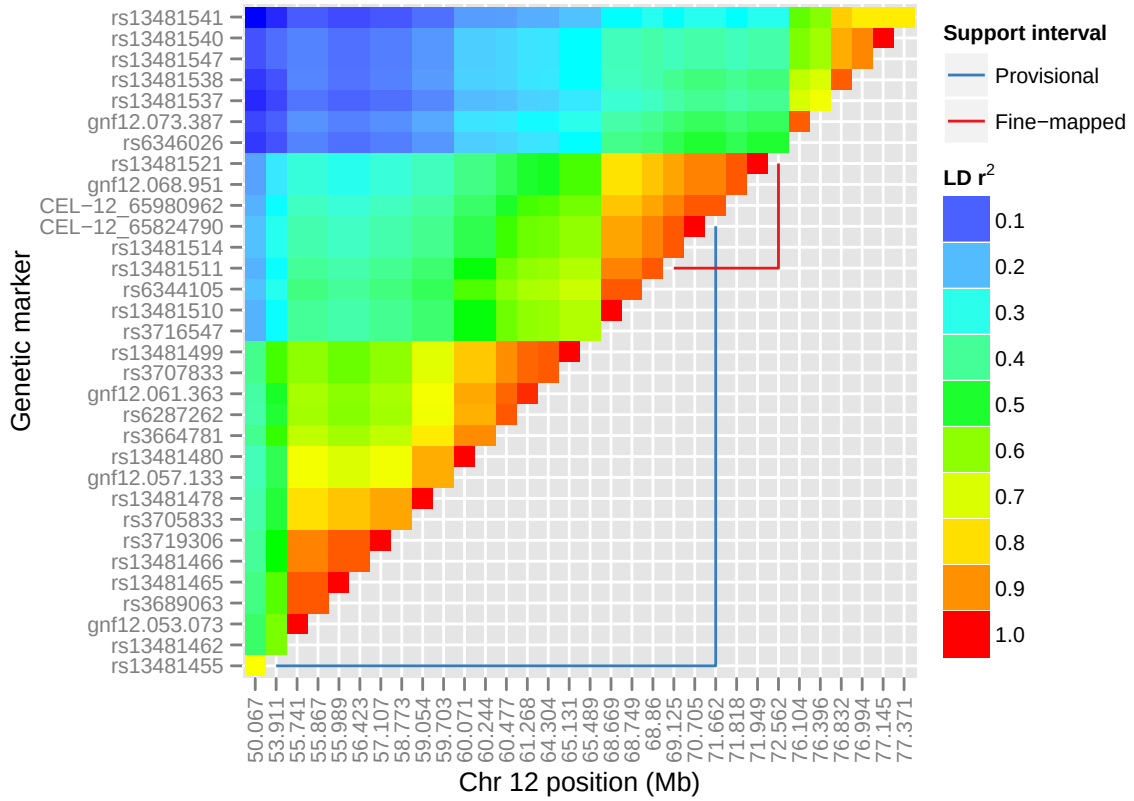

Figure 4.10. Haplotype blocks across *Etanq1* region linkage disequilibrium between pairs of genetic markers near *Etanq1*. Each marker pair is color-coded using the r^2 measure of LD. Red and blue paths define the borders of *Etanq1*'s provisional and fine-mapped support intervals, respectively.

$r^2 \geq 0.7$, then the original *Etanq1* support interval is largely comprised of at least 3 large haplotype blocks. Among the 27 original BXD strains used in the provisional mapping of *Etanq1*, only a handful carry recombination events within this region. It was only with the addition of the novel recombination events provided by the AI BXD strains that differential associations among these haplotype blocks could manifest and the most distal haplotype could emerge as the strongest driver of variation in PDT. These results also indicate that continuing to assay novel strains from the panel of currently available BXD strains is unlikely to narrow *Etanq1*'s support interval any further, as the current support interval almost perfectly encompasses a haplotype block in which the r^2 for all markers is ≥ 0.87.

4.6 Screening for *Etanq1* candidate genes

Given the limited potential for further refinement of the *Etanq1* support interval via genetic mapping, we proceeded to dissect the QTL and prioritize potential candidate QTGs. According to Mouse Genome Informatics website, the fine-mapped support interval for *Etanq1* harbors 41 protein-coding genes or ncRNAs, 66 fewer than the original interval, which substantially reduces the pool of potential candidate genes. Still, 41 genes is not an insubstantial number, particularly in the context of performing molecular validation experiments. We sought to prioritize these positional candidates through a series of integrative analyses that combined phenotypic data from the LD box assay, microarray gene expression data from the PFC, NAc and VMB datasets described in Chapter 2 and genomic sequence data from the B6 and D2 inbred strains.

Table 4.6. Significant *cis* eQTL within *Etanq1*

Gene	Mb	Region	Probe-set	Marker	LOD	p-value
Sos2	70.72	NAc	1443057_at	rs6344105	9.11	1.00E-04
Sos2	70.72	VMB	1443057_at	rs6344105	10.24	0.00E+00
Atp5s	70.83	NAc	1459949_at	rs3716547	5.15	2.90E-03
Map4k5	70.98	NAc	1440059_at	rs6346026	5.17	6.00E-03
Nin	71.11	NAc	1419078_at	rs6344105	4.09	1.30E-02
Trim9	71.35	NAc	1443989_at	rs6344105	4.71	1.40E-02

4.6.1 *cis* eQTL analysis

The microarray expression data was used to identify genes whose expression levels were regulated by local polymorphisms uncovered by the eQTL analysis described in Chapter 3. Only 5 positional candidate genes within *Etanq1*'s support interval were associated with a signifiant *cis* eQTL in either the PFC, NAc or VMB (Table 4.6). All 5 of these genes are clustered within the highly polymorphic region we defined in section 4.5.3 that lies in the center of *Etanq1*'s fine-mapped support interval (Figure 4.9). Although it's not listed in Table 4.6, *Sos2* also has a suggestive *cis* eQTL in PFC that fell just below the significance threshold; still, it stands out as the only gene of the 5 with a strong *cis* eQTL in all three brain regions.

The results from the probe-set SNP analysis described in section 2.6.2 were used here to identify any potentially spurious *cis* eQTL caused by polymorphism hybridization artifacts. Three of the probes within *Map4k5*'s probe-set 1440059_at overlap D2 SNPs; one probe actually overlap two SNPs. Removing the affected probes and repeating the QTL analysis produced only a weak association with local genetic markers, strongly suggesting this was in fact a spurious *cis* eQTL. The probe-set targeting *Sos2*, 1443057_at, contains one probe overlapping a D2 SNP. However, its associated *cis* eQTL in the NAc and VMB were only mildly affected by the removal of this probe and in both regions

Table 4.7. Correlations between expression of *Etanq1*-region genes and PDT

Probe-set	Gene	Mb	*r*	Region	q-value
Nin	1419078_at	71.12	-0.68	NAc	1.59e-03
Sos2	1443057_at	70.72	0.59	NAc	1.71e-02
Trim9	1434595_at	71.37	0.55	NAc	3.31e-02

the association was still significant. As such, unlike *Map4k5*, *Sos2* is still a high priority candidate based on its eQTL results.

4.6.2 Correlation analysis with *Etanq1* candidate genes

Further prioritization was accomplished by correlating each positional candidate's expression with the anxiety behavior of interest, PDT following ethanol, and taking into account the strength of these associations. This correlation analysis was carried out for all three of the profiled brain regions. Figure 4.11 depicts correlational strength between PDT and all genes that reside within the vicinity of *Etanq1*, plotted against their genomic locations. This visualization demonstrates the confounding impact of LD, as the correlational strength gradually begins to rise for all genes the closer there proximity is to the peak of *Etanq1*.

After applying an FDR to correct for the multiple tests conducted within each brain region, we found significant correlations between three positional candidate genes that clearly standout in Figure 4.11. Among these standouts are *Sos2*, *Trim9* and *Nin*. The negative correlation between *Nin*'s expression in the NAc and PDT was particularly robust at (r = -0.67). It's interesting to note that all significant correlations existed between PDT and NAc expression traits.

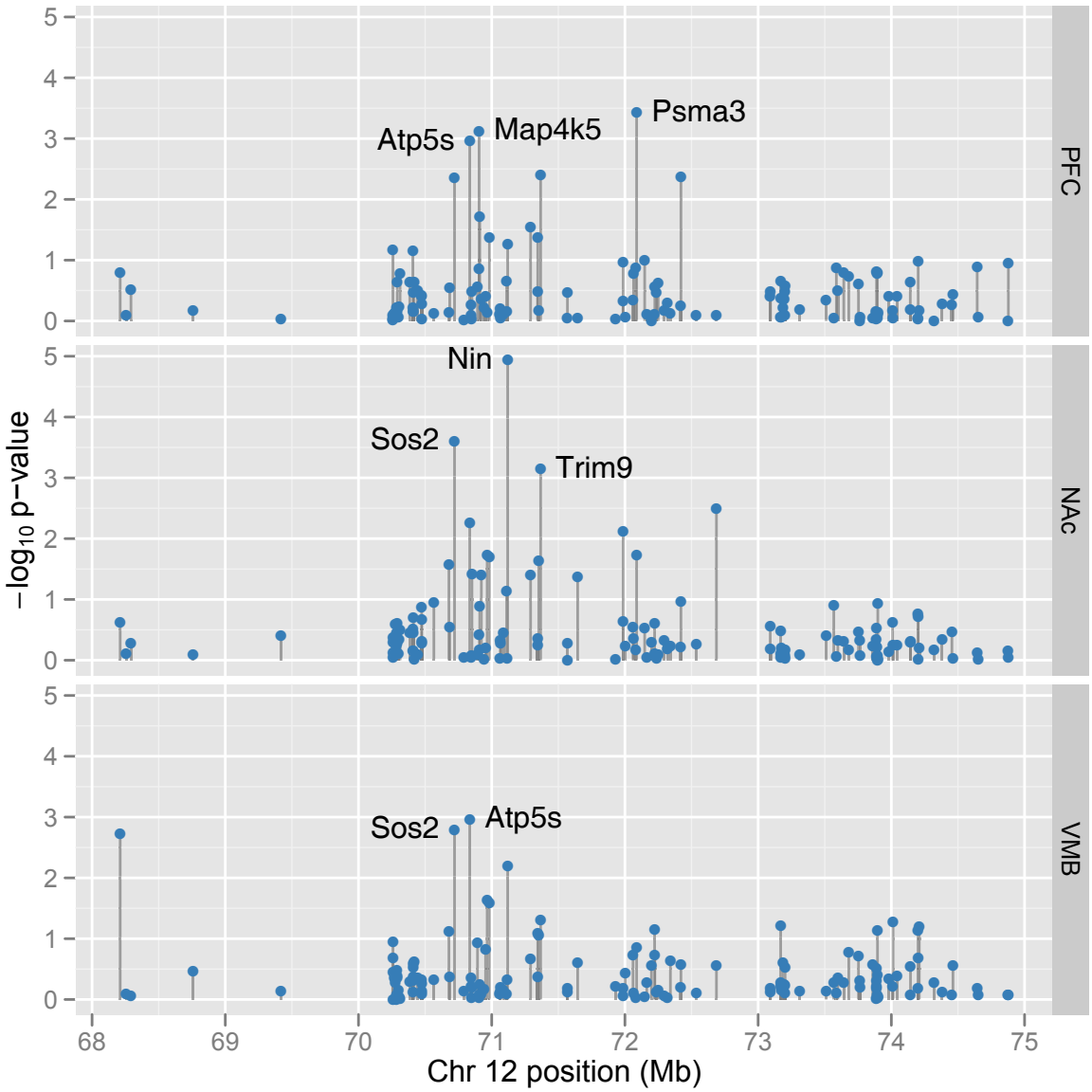

Figure 4.11. Correlations between expression of *Etanq1*-region genes and PDT

Table 4.8. Distribution of transcribed SNPs within *Etanq1*

Gene	Intron	UTR	Exon	Synonymous	Mis-sense
Klhdc1	1	0	0	0	0
Gm71	44	4	2	2	0
Sos2	90	2	1	1	0
L2hgdh	21	1	1	1	0
Atp5s	8	10	1	0	1
Cdkl1	110	1	0	0	0
Map4k5	209	1	3	3	0
Atl1	61	15	1	1	0
Sav1	7	0	0	0	0
Nin	263	38	13	9	4
Pygl	163	1	6	5	1
Trim9	404	0	2	2	0
Tmx1	24	2	1	1	0
Frmd6	7	11	0	0	0
Actr10	3	0	0	0	0
Psma3	5	0	1	1	0
Arid4a	2	2	0	0	0
2700049A03Rik	10	0	1	0	1

All transcribed SNPs within *Etanq1* were analyzed and annotated as members of an intron, an untranslated region or an exon. The right panel indicates whether an exonic SNP causes a synonymous or non-synonymous amino acid (AA) substitution.

4.6.3 SNP analysis of *Etanq1* candidate genes

An analysis of the sequence variation that exists between the B6 and D2 genomes was carried out for all *Etanq1* positional candidates. Unlike the previous strategies for prioritizing candidate genes described in sections 4.6.1 and 4.6.2, the results of this analysis were independent of microarray gene expression data. There are a total of 3,019 SNPs within *Etanq1*'s support interval. Just over half (51.5%) of these SNPs fall within a gene coding region, while only 33 belong to an exon. Collectively, *Trim9* was the most polymorphic gene, harboring over 400 B6D2 SNPs. Though only a handful of

these SNPs fall outside of an intronic region. If we ignore intronic SNPs, *Nin* is the most highly polymorphic gene in the region, comprising by far the largest of untranslated region (UTR) and exonic SNPs. However, our interest was ultimately not in the number of SNPs within a gene, but the potential of each SNP to produce a protein that is functionally polymorphic.

In order to determine whether of these identified coding SNPs are likely to affect protein function, we employed a web-based bioinformatics tool called functional analysis of novel SNPs (FANS), which was developed by Liu et al. (2008) and is accessible at http://fans.ngc.sinica.edu.tw/fans. FANS incorporates several well characterized algorithms for identifying non-synonymous SNPs and predicting whether the AA change will carry functional consequence, including SIFT (Ng and Henikoff, 2003) and PolyPhen (Ramensky et al., 2002), and conveniently integrates the results into a ranking scheme that assesses how "risky" a SNP is to the functional health of the encoded protein. Using FANS to analyze the 33 exonic SNPs within *Etanq1* identified 7 that represent non-synonymous mutations (Table 4.9). Four of these non-synonymous SNPs were located within *Nin*, while the remaining 3 located within *Atp5s*, *Pygl* and a Riken cDNA clone. Of these 7 SNPs, two were identified by FANS as being "high risk," specifically because the altered AAs both fell within a known protein domain. Both high risk SNPs belonged to *Nin* exons.

4.7 Discussion

4.7.1 Candidate gene prioritization

Through the addition of novel BXD strains and variety of integrative genomic analyses we were able to substantially narrow the significant QTL on Chr 12 underlying the

Table 4.9. Etanq1 functional SNP analysis

Gene	Mb	B6 AA	D2 AA	Risk
Atp5s	70,842,781	Val	Ile	Low
Nin	71,144,164	Glu	Lys	Medium
Nin	71,144,373	Arg	Gln	Low
Nin	71,144,376	Ser	Tyr	**High**
Nin	71,144,902	Lys	Glu	**High**
Pygl	71,302,864	Met	Val	Low
2700049a03rik	72,295,279	Asp	Asn	Low

anxiolytic-like response to acute ethanol, originally identified by Putman (2008). The fine-mapped support interval for *Etanq1* now spans a regions less than 4 Mb long and harbors 44 genes. Initially, it appeared that five of these genes were associated with a significant *cis* eQTL in the region. However, the eQTL underlying *Map4k5*'s probed to be a spurious association driven by unaccounted for SNPs within several probe binding regions.

We analyzed the expression patterns for all positional candidate genes in all three brain regions in order to identify genes that were co-expressed with PDT following acute ethanol exposure. Given the important role of the PFC in regulating the amygdala via glutamatergic projections, we hypothesized genes in the PFC would exhibit the strongest relationship with our anxiety-like measurements. However, we observed the strongest correlations existed between PDT and expression levels in the NAc (Figure 4.11). While the NAc is not generally associated with anxiety regulation, several recently published studies indicate that perhaps it should be. Kim et al. (2008) demonstrated that using small interfering RNA (siRNA) to down-regulate *Adcy5* in the NAc, specifically, produced a significant anxiolytic-like response in B6 mice. Furthermore, over-expression of *CREB* in the NAc is able to rescue the enhanced level of anxiety exhibited by rats that have been socially isolated (Barrot et al., 2005).

The results of our correlation analysis also suggested that the NAc plays an important in regulating anxiety-like responses. After correcting for multiple testing, PDT variation was significantly correlated with the expression levels of only three genes—and all three correlation relationships existed in the NAc (Table 4.7). These three genes included *Trim9*, *Sos2* and *Nin*. The strongest association was with ninein (*Nin*)'s expression in NAc, which negatively correlated so that higher levels of *Nin* transcript correspond to a smaller anxiolytic-like response to acute ethanol.

Finally, we also took into account whether positional candidate genes harbored functional polymorphisms that could alter protein function but may have no effect on transcript abundance. A total of 11 genes contained at least one B6/D2 polymorphism with an exon coding region (Table 4.8). Seven of these exonic SNPs represented missense mutations within 4 different genes: *Atp5s*, *Nin*, *Pygl* and an uncharacterized gene, *2700049A03Rik*. *Nin* harbors two SNPs that are predicted to alter in D2 mice the AA of a conserved protein domain that binds DNA and facilitates chromosomal segregation (Table 4.9).

4.7.2 *Ninein* is a strong candidate QTG underlying *Etanq1*

Taken together, these results suggest *Nin* is a strong candidate QTG underlying *Etanq1*. It's difficult to speculate as to the functional role *Nin* might play in mediating ethanol induced anxiolysis, as it primarily characterized as a centrosomal protein that plays a role in microtubule positioning (Stillwell et al., 2004). However, it does share one potentially important link with the genomic analyses of acute ethanol responses performed in Chapter 2 and Chapter 3. Experiments conducted using yeast two-hybrid assays have determined that *Nin* directly interacts and is phosphorylated by *Gsk3β* (Hong et al., 2000; Howng et al., 2004), which was one of top ethanol responsive genes

identified in the PFC (Figure 2.6) and among the most densely interconnected hub within ErGeN3 (Figure 3.7).

Chapter 5

Future directions

Defining complex endophenotypes such as acute ethanol sensitivity in terms of gene networks, rather than the genetic variants that influence them, has the potential to yield information about complex diseases that is more generalizable to humans. Network function, rather than individual gene influences, is likely more conserved evolutionarily. The ethanol-responsive gene-enriched networks defined in Chapter 3 could assist human GWA studies by providing a novel source of functionally related candidate genes. The fact that several of the major ErGeN hub genes have been recently implicated in GWA studies suggests this approach is highly promising. Co-analysis of human GWA studies and ErGeN hub genes may provide bidirectional validation for such genes, even leading to candidates for therapeutic targeting. However, taken out of context, such single genes still do not define the mechanisms underlying cellular, neural network or behavioral responses to ethanol, which remains our chief objective in identifying and dissecting these gene networks.

Direct validation of hub genes, in terms of both gene network regulation and phenotypic responses, are required to fully understand the role of these ethanol-responsive networks in complex behavioral responses. Ongoing studies in the Miles laboratory seek

to adapt and extend this approach, through genetic manipulation of ErGeN hub genes, in order to observe downstream effects on the original ethanol-responsive network as well as the network-associated ethanol behavioral phenotypes. Such validation of network-derived candidates could provide a novel approach to future pharmacotherapies for AUD, directed against regulation of a gene network rather than function of a single protein.

An obvious next step is to extend this research to incorporate the NAc and VMB data in the network analyses. A preliminary analysis of ethanol responsive networks conducted in all three regions using WGCNA (Zhang and Horvath, 2005) produced some interesting results that should be pursued further. For example, the identified gene networks were largely preserved across across all three brain regions (Figure 5.1). That is, comparing the gene constituencies of each network across regions revealed substantial overlap between them Figure 5.2. Furthermore, we found that brown module's eigengene (Zhang and Horvath, 2005) in the NAc was strongly associated with the post-ethanol TLA data generated by (Putman, 2008) in the Miles laboratory ($r = 0.52$, p-value = 9.90E-04). Dissecting this networks could provide novel insight into the molecular mechanism driving this ethanol relevant phenotype.

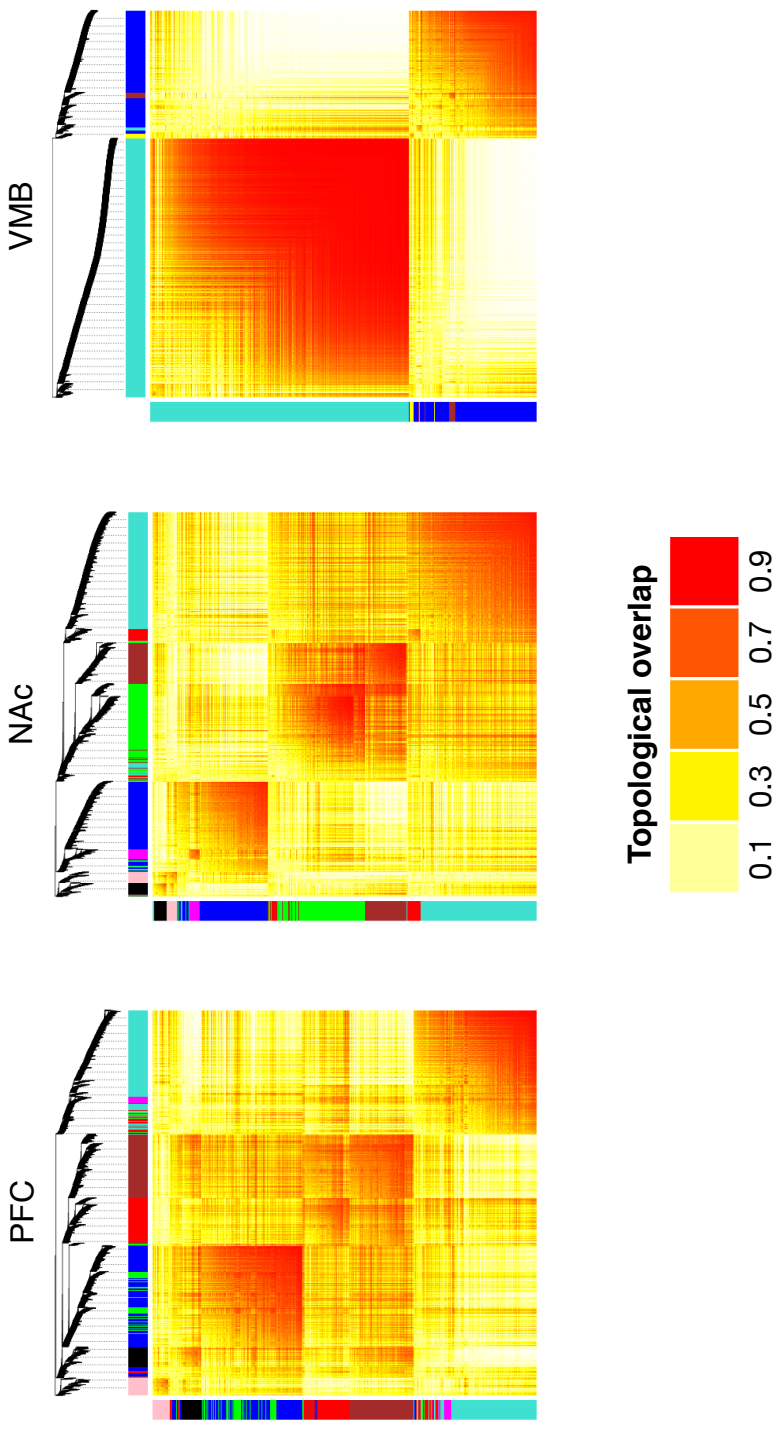

Figure 5.1. Ethanol responsive networks across PFC, NAc and VMB. WGCNA was used to identify gene co-expression modules in the PFC, NAc and VMB. The S-score expression dataset was transformed into a weighted adjacency matrix, providing a measurement of ethanol response similarity for all pairwise gene comparisons. The heatmaps visualize each region's topological overlap matrix, a measure of commonality amongst the network neighborhoods, where warmer colors indicate higher overlap. Individual networks were formed by applying a branch cutting algorithm to the resulting dendrograms above each heatmap, these modules are denoted by the adjacent color bands.

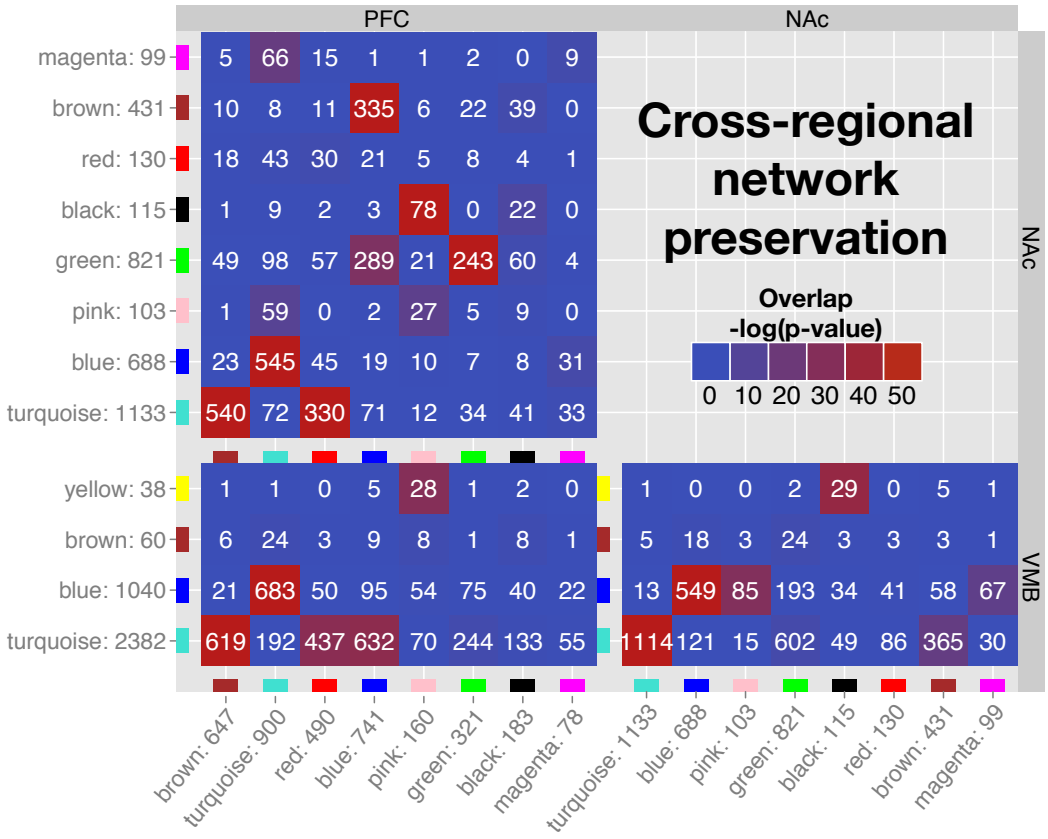

Figure 5.2. Overlap among ethanol responsive networks across PFC, NAc and VMB. The contingency tables above indicate the number of probe-sets in common between the corresponding row and column modules. The total probe-set counts for each module are provided in the axis labels. Fisher's exact test was used to to determine the statistical significance of overlapping modules, -log(p-values) from this analysis were used to color code the table.

Future work will focus on incorporating more advanced statistical methods that make it possible to refine the connections within a gene network and infer causal relationships among genetic variants, high-throughput molecular data and complex phenotypes. An excellent example of how effective such methods can be is provided by Schadt et al. (2005). Using the same liver expression dataset described in section 1.5.3, and a novel network construction technique called likelihood-based causality model selection (LCMS), the investigators first identified all QTLs associated with a classical

phenotype and then winnowed the list of potentially associated gene-expression traits on the basis of their correlation or eQTL overlap with the phenotype of interest, FPM. Candidate genes then were ranked by applying using LCMS, which uses the eQTL data to establish causal relationships between genetic loci and transcripts, as well as between transcripts and phenotypes, and finally identifies a model that best fits the data.

By ranking genes according to their performance in these models, the investigators identified several novel obesity candidate genes as well as uncovered additional support for the involvement of a gene called *Hsd11b1* that previously had been implicated in obesity risk (Rask et al., 2002). Because this gene seemed to be relevant to the phenotype they were investigating, the researchers then sought to reconstruct the gene network in which *Hsd11b1* participates by performing the LCMS procedure with *Hsd11b1* as the trait of interest. The resulting network was able to successfully predict genes that would be affected by inhibition of *Hsd11b1*. This progression from phenotype to gene network to candidate gene and back to a gene network is a striking example of the promise that combining genetical genomics and gene-network analysis provides for understanding complex traits such as alcoholism.

Bibliography

Abiola, O., Angel, J. M., Avner, P., Bachmanov, A. A., Belknap, J. K., Bennett, B., Blankenhorn, E. P., Blizard, D. A., Bolivar, V., Brockmann, G. A., Buck, K. J., Bureau, J.-F., Casley, W. L., Chesler, E. J., Cheverud, J. M., Churchill, G. A., Cook, M., Crabbe, J. C., Crusio, W. E., Darvasi, A., de Haan, G., Dermant, P., Doerge, R. W., Elliot, R. W., Farber, C. R., Flaherty, L., Flint, J., Gershenfeld, H., Gibson, J. P., Gu, J., Gu, W., Himmelbauer, H., Hitzemann, R., Hsu, H.-C., Hunter, K., Iraqi, F. F., Jansen, R. C., Johnson, T. E., Jones, B. C., Kempermann, G., Lammert, F., Lu, L., Manly, K. F., Matthews, D. B., Medrano, J. F., Mehrabian, M., Mittlemann, G., Mock, B. A., Mogil, J. S., Montagutelli, X., Morahan, G., Mountz, J. D., Nagase, H., Nowakowski, R. S., O'Hara, B. F., Osadchuk, A. V., Paigen, B., Palmer, A. A., Peirce, J. L., Pomp, D., Rosemann, M., Rosen, G. D., Schalkwyk, L. C., Seltzer, Z., Settle, S., Shimomura, K., Shou, S., Sikela, J. M., Siracusa, L. D., Spearow, J. L., Teuscher, C., Threadgill, D. W., Toth, L. A., Toye, A. A., Vadasz, C., Van Zant, G., Wakeland, E., Williams, R. W., Zhang, H.-G., Zou, F., and Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet*, 4(11):911–6. 71, 99

Abu-Khzam, F. N., Langston, M. A., Shanbhag, P., and Symons, C. T. (2006). Scalable parallel algorithms for FPT problems. *Algorithmica*, 45(3):269–284. 62

Affymetrix (2002). Statistical algorithms description document. Technical report, Santa Clara, CA. 30

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.-P., and Jansen, R. C. (2007). Sequence polymorphisms cause many false cis eqtls. *PLoS One*, 2(7):e622. 73

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–6. 38

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–8. 3

Analytics, R. (2011). *foreach: Foreach looping construct for R*. R package version 1.3.2. 158

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9. 11

Bäckström, P. and Hyytiä, P. (2005). Suppression of alcohol self-administration and cue-induced reinstatement of alcohol seeking by the mglu2/3 receptor agonist LY379268 and the mGlu8 receptor agonist (S)-3,4-DCPG. *Eur J Pharmacol*, 528(1-3):110–8. 90

Bailey, D. W. (1971). Recombinant-inbred strains. an aid to finding identity, linkage, and function of histocompatibility and other genes. *Transplantation*, 11(3):325–7. 5

Baldwin, N. E., Chesler, E. J., Kirov, S., Langston, M. A., Snoddy, J. R., Williams, R. W., and Zhang, B. (2005). Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks. *J Biomed Biotechnol*, 2005(2):172–80. 62

Barabasi and Albert (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–12. 61

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.-P., de Vos, M., Vermeire, S., Louis, E., Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet*, 40(8):955–62. 3

Barrot, M., Wallace, D. L., Bolaños, C. A., Graham, D. L., Perrotti, L. I., Neve, R. L., Chambliss, H., Yin, J. C., and Nestler, E. J. (2005). Regulation of anxiety and initiation of sexual behavior by creb in the nucleus accumbens. *Proc Natl Acad Sci U S A*, 102(23):8357–62. 120

Becker, H. C. and Lopez, M. F. (2004). Increased ethanol drinking after repeated chronic ethanol exposure and withdrawal experience in C57BL/6 mice. *Alcohol Clin Exp Res*, 28(12):1829–38. 53

Belknap, J. K. and Atkins, A. L. (2001). The replicability of QTLs for murine alcohol preference drinking behavior across eight independent studies. *Mamm Genome*, 12(12):893–9. 17

Belknap, J. K., Metten, P., Helms, M. L., O'Toole, L. A., Angeli-Gade, S., Crabbe, J. C., and Phillips, T. J. (1993). Quantitative trait loci (qtl) applications to substances of abuse: physical dependence studies with nitrous oxide and ethanol in bxd mice. *Behav Genet*, 23(2):213–22. 54

Belknap, J. K., Mogil, J. S., Helms, M. L., Richards, S. P., O'Toole, L. A., Bergeson, S. E., and Buck, K. J. (1995). Localization to chromosome 10 of a locus influencing morphine analgesia in crosses derived from C57BL/6 and DBA/2 strains. *Life Sci*, 57(10):PL117–24. 88

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., and Marron, J. S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–14. 38

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1):289–300. 41

Bennett, B., Beeson, M., Gordon, L., Carosone-Link, P., and Johnson, T. E. (2002). Genetic dissection of quantitative trait loci specifying sedative/hypnotic sensitivity to ethanol: mapping with interval-specific congenic recombinant lines. *Alcohol Clin Exp Res*, 26(11):1615–24. 84

Bennett, B., Downing, C., Parker, C., and Johnson, T. E. (2006). Mouse genetic models in alcohol research. *Trends Genet*, 22(7):367–74. 8

Boehm, 2nd, S. L., Reed, C. L., McKinnon, C. S., and Phillips, T. J. (2002). Shared genes influence sensitivity to the effects of ethanol on locomotor and anxiety-like behaviors, and the stress axis. *Psychopharmacology (Berl)*, 161(1):54–63. 93, 94

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–31. 2

Breese, G. R., Criswell, H. E., Carta, M., Dodson, P. D., Hanchar, H. J., Khisti, R. T., Mameli, M., Ming, Z., Morrow, A. L., Olsen, R. W., Otis, T. S., Parsons, L. H., Penland, S. N., Roberto, M., Siggins, G. R., Valenzuela, C. F., and Wallner, M. (2006). Basis of the gabamimetic profile of ethanol. *Alcohol Clin Exp Res*, 30(4):731–44. 55

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–5. 14

Broman, K. W. and Sen, S. (2009). *A guide to QTL mapping with R/qtl*. Statistics for biology and health. Springer, Dordrecht. 4

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–90. 71

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21(1 Suppl):33–7. 25

Buck, K., Metten, P., Belknap, J., and Crabbe, J. (1999). Quantitative trait loci affecting risk for pentobarbital withdrawal map near alcohol withdrawal loci on mouse chromosomes 1, 4, and 11. *Mamm Genome*, 10(5):431–7. 6

Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29. 61

Cappell, H. and Herman, C. P. (1972). Alcohol and tension reduction. a review. *Q J Stud Alcohol*, 33(1):33–64. 93

Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–50. 61

Chandler, L. J., Carpenter-Hyland, E., Hendricson, A. W., Maldve, R. E., Morrisett, R. A., Zhou, F. C., Sari, Y., Bell, R., and Szumlinski, K. K. (2006). Structural and functional modifications in glutamateric synapses following prolonged ethanol exposure. *Alcohol Clin Exp Res*, 30(2):368–76. 90

Chandra, S., Fornai, F., Kwon, H.-B., Yazdani, U., Atasoy, D., Liu, X., Hammer, R. E., Battaglia, G., German, D. C., Castillo, P. E., and Südhof, T. C. (2004). Double-knockout mice for alpha- and beta-synucleins: effect on synaptic functions. *Proc Natl Acad Sci U S A*, 101(41):14966–71. 82

Chen, G., Bower, K. A., Xu, M., Ding, M., Shi, X., Ke, Z.-J., and Luo, J. (2009a). Cyanidin-3-glucoside reverses ethanol-induced inhibition of neurite outgrowth: role of glycogen synthase kinase 3 beta. *Neurotox Res*, 15(4):321–31. 79

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009b). Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*, 37(Web Server issue):W305–11. 51

Chesler, E. J. and Langston, M. A. (2005). Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data. *Proceedings, RECOMB Satellite Workshop on Systems Biology and Regulatory Genomics*, page 17. 62

Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., Threadgill, D. W., Manly, K. F., and Williams, R. W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*, 37(3):233–42. 14, 60

Chesler, E. J., Wang, J., Lu, L., Qu, Y., Manly, K. F., and Williams, R. W. (2003). Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. *Neuroinformatics*, 1(4):343–57. 14

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat Genet*, 32 Suppl:490–5. 25

Cook, T. A. R., Luczak, S. E., Shea, S. H., Ehlers, C. L., Carr, L. G., and Wall, T. L. (2005). Associations of aldh2 and adh1b genotypes with response to alcohol in asian americans. *J Stud Alcohol*, 66(2):196–204. 3

Cooper, T. F., Morby, A. P., Gunn, A., and Schneider, D. (2006). Effect of random and hub gene disruptions on environmental and mutational robustness in escherichia coli. *BMC Genomics*, 7:237. 13

Crabbe, J. C., Kosobud, A., Young, E. R., and Janowsky, J. S. (1983). Polygenic and single-gene determination of responses to ethanol in BXD/Ty recombinant inbred mouse strains. *Neurobehav Toxicol Teratol*, 5(2):181–7. 88, 89

Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science*, 284(5420):1670–2. 103

Crawley, J. and Goodwin, F. K. (1980). Preliminary report of a simple animal behavior model for the anxiolytic effects of benzodiazepines. *Pharmacol Biochem Behav*, 13(2):167–70. 94

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. 64

Cunningham, C. L. (1995). Localization of genes influencing ethanol-induced conditioned place preference and locomotor activity in BXD recombinant inbred mice. *Psychopharmacology (Berl)*, 120(1):28–41. 89

Damerval, C., Maurice, A., Josse, J. M., and de Vienne, D. (1994). Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics*, 137(1):289–301. 13

Daniels, G. M. and Buck, K. J. (2002). Expression profiling identifies strain-specific changes associated with ethanol withdrawal in mice. *Genes Brain Behav*, 1(1):35–45. 10

Davies, A. G., Pierce-Shimomura, J. T., Kim, H., VanHoven, M. K., Thiele, T. R., Bonci, A., Bargmann, C. I., and McIntire, S. L. (2003). A central role of the BK potassium channel in behavioral responses to ethanol in C. elegans. *Cell*, 115(6):655–66. 55, 90

DeFries, J. C., Wilson, J. R., Erwin, V. G., and Petersen, D. R. (1989). Ls x ss recombinant inbred strains of mice: initial characterization. *Alcohol Clin Exp Res*, 13(2):196–200. 5

Dick, D. M., Bierut, L., Hinrichs, A., Fox, L., Bucholz, K. K., Kramer, J., Kuperman, S., Hesselbrock, V., Schuckit, M., Almasy, L., Tischfield, J., Porjesz, B., Begleiter, H., Nurnberger, Jr, J., Xuei, X., Edenberg, H. J., and Foroud, T. (2006). The role of gabra2 in risk for conduct disorder and alcohol and drug dependence across developmental stages. *Behav Genet*, 36(4):577–90. 3

Do, K.-A., Müller, P., and Vannucci, M. (2006). *Bayesian inference for gene expression and proteomics*. Cambridge University Press, Cambridge. 26

Dobrin, R., Zhu, J., Molony, C., Argman, C., Parrish, M. L., Carlson, S., Allan, M. F., Pomp, D., and Schadt, E. E. (2009). Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol*, 10(5):R55. 13

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet*, 3(1):43–52. 3

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–94. 71

Dong, J. and Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst Biol*, 1:24. 12

Dopico, A. M., Lemos, J. R., and Treistman, S. N. (1996). Ethanol increases the activity of large conductance, Ca(2+)-activated K+ channels in isolated neurohypophysial terminals. *Mol Pharmacol*, 49(1):40–8. 79, 90

Doss, S., Schadt, E. E., Drake, T. A., and Lusis, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Res*, 15(5):681–91. 57, 73

Eblen, J. D., Jay, J. J., Zhang, Y., Benson, M., Perkins, A. D., Saxton, A. M., Voy, B. H., Chesler, E. J., and Langston, M. A. (2011). A Systematic Comparison of Genome Scale Clustering Algorithms. *International Symposium on Bioinformatics Research and Applications*. 63

Edenberg, H. J., Dick, D. M., Xuei, X., Tian, H., Almasy, L., Bauer, L. O., Crowe, R. R., Goate, A., Hesselbrock, V., Jones, K., Kwon, J., Li, T.-K., Nurnberger, Jr, J. I., O'Connor, S. J., Reich, T., Rice, J., Schuckit, M. A., Porjesz, B., Foroud, T., and Begleiter, H. (2004). Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor,

are associated with alcohol dependence and with brain oscillations. *Am J Hum Genet*, 74(4):705–14. 3, 55

Edenberg, H. J., Koller, D. L., Xuei, X., Wetherill, L., McClintick, J. N., Almasy, L., Bierut, L. J., Bucholz, K. K., Goate, A., Aliev, F., Dick, D., Hesselbrock, V., Hinrichs, A., Kramer, J., Kuperman, S., Nurnberger, Jr, J. I., Rice, J. P., Schuckit, M. A., Taylor, R., Todd Webb, B., Tischfield, J. A., Porjesz, B., and Foroud, T. (2010). Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol Clin Exp Res*, 34(5):840–52. 90

Egan, M. F., Straub, R. E., Goldberg, T. E., Yakub, I., Callicott, J. H., Hariri, A. R., Mattay, V. S., Bertolino, A., Hyde, T. M., Shannon-Weickert, C., Akil, M., Crook, J., Vakkalanka, R. K., Balkissoon, R., Gibbs, R. A., Kleinman, J. E., and Weinberger, D. R. (2004). Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. *Proc Natl Acad Sci U S A*, 101(34):12604–9. 90

Elliott, R. C., Miles, M. F., and Lowenstein, D. H. (2003). Overlapping microarray profiles of dentate gyrus gene expression during development- and epilepsy-associated neurogenesis and axon outgrowth. *J Neurosci*, 23(6):2218–27. 34

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K. P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H. G., Stefansson, T., Leifsson, B. G., Thorsteinsdottir, U., Lamb, J. R., Gulcher, J. R., Reitman, M. L., Kong, A., Schadt, E. E., and Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–8. 18

Fehr, C., Shirley, R. L., Belknap, J. K., Crabbe, J. C., and Buck, K. J. (2002). Congenic mapping of alcohol and pentobarbital withdrawal liability loci to a <1 centimorgan interval of murine chromosome 4: identification of mpdz as a candidate gene. *J Neurosci*, 22(9):3730–8. 6

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh. 40

Flint, J., Valdar, W., Shifman, S., and Mott, R. (2005). Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Genet*, 6(4):271–86. 6, 95

Foroud, T., Edenberg, H. J., Goate, A., Rice, J., Flury, L., Koller, D. L., Bierut, L. J., Conneally, P. M., Nurnberger, J. I., Bucholz, K. K., Li, T. K., Hesselbrock, V., Crowe, R., Schuckit, M., Porjesz, B., Begleiter, H., and Reich, T. (2000). Alcoholism susceptibility loci: confirmation studies in a replicate sample and further mapping. *Alcohol Clin Exp Res*, 24(7):933–45. 1

French, R. L. and Heberlein, U. (2009). Glycogen synthase kinase-3/shaggy mediates ethanol-induced excitotoxic cell death of drosophila olfactory neurons. *Proc Natl Acad Sci U S A*, 106(49):20924–9. 79

Garlow, S. J., Boone, E., Li, W., Owens, M. J., and Nemeroff, C. B. (2005). Genetic analysis of the hypothalamic corticotropin-releasing factor system. *Endocrinology*, 146(5):2362–8. 82

Gass, J. T. and Olive, M. F. (2008). Glutamatergic substrates of drug addiction and alcoholism. *Biochem Pharmacol*, 75(1):218–65. 79, 90

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80. 20, 39

Gilad, Y., Rifkin, S. A., Bertone, P., Gerstein, M., and White, K. P. (2005). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res*, 15(5):674–80. 57

Gill, K. J. and Boyle, A. E. (2003). Confirmation of quantitative trait loci for cocaine-induced activation in the AcB/BcA series of recombinant congenic strains. *Pharmacogenetics*, 13(6):329–38. 82

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–6. 12

Grahame, N. J., Li, T. K., and Lumeng, L. (1999). Limited access alcohol drinking in high- and low-alcohol preferring selected lines of mice. *Alcohol Clin Exp Res*, 23(6):1015–22. 9

Grieve, S. J. and Littleton, J. M. (1979). Age and strain differences in the rat of development of functional tolerance to ethanol by mice. *J Pharm Pharmacol*, 31(10):696–700. 4, 54

Grisel, J. E. (2000). Quantitative trait locus analysis. *Alcohol Res Health*, 24(3):169–74. 3

Grisel, J. E., Metten, P., Wenger, C. D., Merrill, C. M., and Crabbe, J. C. (2002). Mapping of quantitative trait loci underlying ethanol metabolism in bxd recombinant inbred mouse strains. *Alcohol Clin Exp Res*, 26(5):610–6. 89

Haiman, C. A., Le Marchand, L., Yamamato, J., Stram, D. O., Sheng, X., Kolonel, L. N., Wu, A. H., Reich, D., and Henderson, B. E. (2007a). A common genetic risk factor for colorectal and prostate cancer. *Nat Genet*, 39(8):954–6. 3

Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G. J., Greenway, S. C., Stram, D. O., Le Marchand, L., Kolonel, L. N., Frasco, M., Wong, D., Pooler, L. C., Ardlie, K., Oakley-Girvan, I., Whittemore, A. S., Cooney, K. A., John, E. M., Ingles, S. A., Altshuler, D., Henderson, B. E., and Reich, D. (2007b). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet*, 39(5):638–44. 3

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity (Edinb)*, 69(4):315–24. 71

Hardin, J. and Wilson, J. (2009). A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, 10(3):446–50. 28

Heath, A. C., Bucholz, K. K., Madden, P. A., Dinwiddie, S. H., Slutske, W. S., Bierut, L. J., Statham, D. J., Dunne, M. P., Whitfield, J. B., and Martin, N. G. (1997). Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychol Med*, 27(6):1381–96. 1

Hill, S. Y., Shen, S., Zezza, N., Hoffman, E. K., Perlin, M., and Allan, W. (2004). A genome wide search for alcoholism susceptibility genes. *Am J Med Genet B Neuropsychiatr Genet*, 128B(1):102–13. 1

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108. 3

Hirschhorn, J. N., Lindgren, C. M., Daly, M. J., Kirby, A., Schaffner, S. F., Burtt, N. P., Altshuler, D., Parker, A., Rioux, J. D., Platko, J., Gaudet, D., Hudson, T. J., Groop, L. C., and Lander, E. S. (2001). Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *Am J Hum Genet*, 69(1):106–16. 2

Hong, Y. R., Chen, C. H., Chang, J. H., Wang, S., Sy, W. D., Chou, C. K., and Howng, S. L. (2000). Cloning and characterization of a novel human ninein protein that interacts with the glycogen synthase kinase 3beta. *Biochim Biophys Acta*, 1492(2-3):513–6. 121

Horishita, T. and Harris, R. A. (2008). n-alcohols inhibit voltage-gated na+ channels expressed in xenopus oocytes. *J Pharmacol Exp Ther*, 326(1):270–7. 91

Howng, S.-L., Hsu, H.-C., Cheng, T.-S., Lee, Y.-L., Chang, L.-K., Lu, P.-J., and Hong, Y.-R. (2004). A novel ninein-interaction protein, cgi-99, blocks ninein phosphorylation by gsk3beta and is highly expressed in brain tumors. *FEBS Lett*, 566(1-3):162–8. 121

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64. 25, 28, 30, 34, 39

Jansen, R. C. and Nap, J. P. (2001). Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–91. 14

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4. 61

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–27. 38

Jones, B. C., Tarantino, L. M., Rodriguez, L. A., Reed, C. L., McClearn, G. E., Plomin, R., and Erwin, V. G. (1999). Quantitative-trait loci analysis of cocaine-related behaviours and neurochemistry. *Pharmacogenetics*, 9(5):607–17. 88

Junker, B. H. and Schreiber, F. (2008). *Analysis of biological networks*. Wiley-Interscience, Hoboken, N.J. 10

Kalivas, P. W., Volkow, N., and Seamans, J. (2005). Unmanageable motivation in addiction: a pathology in prefrontal-accumbens glutamate transmission. *Neuron*, 45(5):647–50. 63

Kao, W.-T., Wang, Y., Kleinman, J. E., Lipska, B. K., Hyde, T. M., Weinberger, D. R., and Law, A. J. (2010). Common genetic variation in Neuregulin 3 (NRG3) influences risk for schizophrenia and impacts NRG3 expression in human brain. *Proc Natl Acad Sci U S A*, 107(35):15619–24. 90

Kapfhamer, D., Bettinger, J. C., Davies, A. G., Eastman, C. L., Smail, E. A., Heberlein, U., and McIntire, S. L. (2008). Loss of RAB-3/A in caenorhabditis elegans and the mouse affects behavioral response to ethanol. *Genes Brain Behav*, 7(6):669–76. 86

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E. S., Ordovas, J. M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M.-R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D. M., and Orho-Melander, M. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet*, 40(2):189–97. 18

Kendler, K. S., Kalsi, G., Holmans, P. A., Sanders, A. R., Aggen, S. H., Dick, D. M., Aliev, F., Shi, J., Levinson, D. F., and Gejman, P. V. (2011). Genomewide association analysis of symptoms of alcohol dependence in the molecular genetics of schizophrenia (MGS2) control sample. *Alcohol Clin Exp Res*, 35(5):963–75. 90

Kendler, K. S., Neale, M. C., Heath, A. C., Kessler, R. C., and Eaves, L. J. (1994). A twin-family study of alcoholism in women. *Am J Psychiatry*, 151(5):707–15. 1

Kennedy, R. E., Archer, K. J., and Miles, M. F. (2006a). Empirical validation of the S-Score algorithm in the analysis of gene expression data. *BMC Bioinformatics*, 7:154. 34

Kennedy, R. E., Kerns, R. T., Kong, X., Archer, K. J., and Miles, M. F. (2006b). SScore: an R package for detecting differential gene expression without gene expression summaries. *Bioinformatics*, 22(10):1272–4. 40

Kerns, R. T., Ravindranathan, A., Hassan, S., Cage, M. P., York, T., Sikela, J. M., Williams, R. W., and Miles, M. F. (2005). Ethanol-responsive brain region expression networks: implications for behavioral responses to acute ethanol in DBA/2J versus C57BL/6J mice. *J Neurosci*, 25(9):2255–66. 9, 19, 36, 41, 43, 52, 55, 63

Kerns, R. T., Zhang, L., and Miles, M. F. (2003). Application of the S-score algorithm for analysis of oligonucleotide microarrays. *Methods*, 31(4):274–81. 32, 33, 34

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., and Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the united states. results from the national comorbidity survey. *Arch Gen Psychiatry*, 51(1):8–19. 93

Kholodenko, B. N., Demin, O. V., Moehren, G., and Hoek, J. B. (1999). Quantification of short term signaling by the epidermal growth factor receptor. *J Biol Chem*, 274(42):30169–81. 48

Kim, K.-S., Lee, K.-W., Baek, I.-S., Lim, C.-M., Krishnan, V., Lee, J.-K., Nestler, E. J., and Han, P.-L. (2008). Adenylyl cyclase-5 activity in the nucleus accumbens regulates anxiety-related behavior. *J Neurochem*, 107(1):105–15. 120

Kirstein, S. L., Davidson, K. L., Ehringer, M. A., Sikela, J. M., Erwin, V. G., and Tabakoff, B. (2002). Quantitative trait loci affecting initial sensitivity and acute functional tolerance to ethanol-induced ataxia and brain camp signaling in bxd recombinant inbred mice. *J Pharmacol Exp Ther*, 302(3):1238–45. 17

Koob, G. F. (2003). Alcoholism: allostasis and beyond. *Alcohol Clin Exp Res*, 27(2):232–43. 53

Kuhn, K., Baker, S. C., Chudin, E., Lieu, M.-H., Oeser, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. K., and Chee, M. S. (2004). A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res*, 14(11):2347–56. 23

Kuo, P.-H., Kalsi, G., Prescott, C. A., Hodgkinson, C. A., Goldman, D., van den Oord, E. J., Alexander, J., Jiang, C., Sullivan, P. F., Patterson, D. G., Walsh, D., Kendler, K. S., and Riley, B. P. (2008). Association of adh and aldh genes with alcohol dependence in the irish affected sib pair study of alcohol dependence (iaspsad) sample. *Alcohol Clin Exp Res*, 32(5):785–95. 3

LaBuda, C. J. and Fuchs, P. N. (2000). Aspirin attenuates the anxiolytic actions of ethanol. *Alcohol*, 21(3):287–90. 93

LaBuda, C. J. and Fuchs, P. N. (2001). The anxiolytic effect of acute ethanol or diazepam exposure is unaltered in mu-opioid receptor knockout mice. *Brain Res Bull*, 55(6):755–60. 93

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–35. 38

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559. 62

Langston, M. A., Perkins, A. D., Saxton, A. M., Scharff, J. A., and Voy, B. H. (2008). Innovative computational methods for transcriptomic data analysis: A case study in the use of FPT for practical algorithm design and implementation. *The Computer Journal*, 51(1):26–38. 62

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., and Fraser, A. G. (2006). Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, 38(8):896–903. 13

Lennon, G. G. and Lehrach, H. (1991). Hybridization analyses of arrayed cDNA libraries. *Trends Genet*, 7(10):314–7. 22

Lessov, C. N., Palmer, A. A., Quick, E. A., and Phillips, T. J. (2001). Voluntary ethanol drinking in C57BL/6J and DBA/2J mice before and after sensitization to the locomotor stimulant effects of ethanol. *Psychopharmacology (Berl)*, 155(1):91–9. 94

Lewohl, J. M., Wang, L., Miles, M. F., Zhang, L., Dodd, P. R., and Harris, R. A. (2000). Gene expression in human alcoholism: microarray analysis of frontal cortex. *Alcohol Clin Exp Res*, 24(12):1873–82. 8, 9

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–6. 30

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–4. 23

Liu, C.-K., Chen, Y.-H., Tang, C.-Y., Chang, S.-C., Lin, Y.-J., Tsai, M.-F., Chen, Y.-T., and Yao, A. (2008). Functional analysis of novel snps and mutations in human and mouse genomes. *BMC Bioinformatics*, 9 Suppl 12:S10. 119

Liu, J., Lewohl, J. M., Harris, R. A., Iyer, V. R., Dodd, P. R., Randall, P. K., and Mayfield, R. D. (2006). Patterns of gene expression in the frontal cortex discriminate alcoholic from nonalcoholic individuals. *Neuropsychopharmacology*, 31(7):1574–82. 63

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80. 23

Manichaikul, A., Dupuis, J., Sen, S., and Broman, K. W. (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics*, 174(1):481–9. 71, 99

Markel, P. D., Fulker, D. W., Bennett, B., Corley, R. P., DeFries, J. C., Erwin, V. G., and Johnson, T. E. (1996). Quantitative trait loci for ethanol sensitivity in the ls x ss recombinant inbred strains: interval mapping. *Behav Genet*, 26(4):447–58. 84

Martin, M. V., Dong, H., Vallera, D., Lee, D., Lu, L., Williams, R. W., Rosen, G. D., Cheverud, J. M., and Csernansky, J. G. (2006). Independent quantitative trait loci influence ventral and dorsal hippocampal volume in recombinant inbred strains of mice. *Genes Brain Behav*, 5(8):614–23. 88

Mayfield, R. D., Lewohl, J. M., Dodd, P. R., Herlihy, A., Liu, J., and Harris, R. A. (2002). Patterns of gene expression are altered in the frontal and motor cortices of human alcoholics. *J Neurochem*, 81(4):802–13. 8, 9

McClintick, J. N. and Edenberg, H. J. (2006). Effects of filtering by present call on analysis of microarray experiments. *BMC Bioinformatics*, 7:49. 39

Metten, P. and Crabbe, J. (1994). Common genetic determinants of severity of acute withdrawal from ethanol, pentobarbital and diazepam in inbred mice. *Behav Pharmacol*, 5(4 And 5):533–547. 4, 54

Metten, P., Phillips, T. J., Crabbe, J. C., Tarantino, L. M., McClearn, G. E., Plomin, R., Erwin, V. G., and Belknap, J. K. (1998). High genetic susceptibility to ethanol withdrawal predicts low ethanol consumption. *Mamm Genome*, 9(12):983–90. 92

Miles, M. F. (2001). Microarrays: lost in a storm of data? *Nat Rev Neurosci*, 2(6):441–3. 25

Miller, J. A., Horvath, S., and Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proc Natl Acad Sci U S A*, 107(28):12698–703. 7

Moghaddam, B. and Adams, B. W. (1998). Reversal of phencyclidine effects by a group ii metabotropic glutamate receptor agonist in rats. *Science*, 281(5381):1349–52. 90

Morar, B., Dragović, M., Waters, F. A. V., Chandler, D., Kalaydjieva, L., and Jablensky, A. (2011). Neuregulin 3 (NRG3) as a susceptibility gene in a schizophrenia subtype with florid delusions and relatively spared cognition. *Mol Psychiatry*, 16(8):860–6. 90

Mozhui, K., Ciobanu, D. C., Schikorski, T., Wang, X., Lu, L., and Williams, R. W. (2008). Dissection of a QTL hotspot on mouse distal chromosome 1 that modulates neurobehavioral phenotypes and gene expression. *PLoS Genet*, 4(11):e1000260. 15, 99

Mulligan, M. K., Ponomarev, I., Hitzemann, R. J., Belknap, J. K., Tabakoff, B., Harris, R. A., Crabbe, J. C., Blednov, Y. A., Grahame, N. J., Phillips, T. J., Finn, D. A., Hoffman, P. L., Iyer, V. R., Koob, G. F., and Bergeson, S. E. (2006). Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc Natl Acad Sci U S A*, 103(16):6368–73. 9, 53, 55

Mulligan, M. K., Rhodes, J. S., Crabbe, J. C., Mayfield, R. D., Adron Harris, R., and Ponomarev, I. (2011). Molecular profiles of drinking alcohol to intoxication in c57bl/6j mice. *Alcohol Clin Exp Res*, 35(4):659–70. 10, 61

Nestoros, J. N. (1980). Ethanol specifically potentiates gaba-mediated neurotransmission in feline cerebral cortex. *Science*, 209(4457):708–10. 55

Newlin, D. B. and Thomson, J. B. (1990). Alcohol challenge with sons of alcoholics: a critical review and analysis. *Psychol Bull*, 108(3):383–402. 93

Ng, P. C. and Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–4. 119

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*, 6(4):e1000888. 18

Nielsen, J. and Oliver, S. (2005). The next wave in metabolome analysis. *Trends Biotechnol*, 23(11):544–6. 7

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nat Neurosci*, 11(11):1271–82. 7

Peirce, J. L., Lu, L., Gu, J., Silver, L. M., and Williams, R. W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet*, 5:7. 5, 99, 108

Peters, L. L., Robledo, R. F., Bult, C. J., Churchill, G. A., Paigen, B. J., and Svenson, K. L. (2007). The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet*, 8(1):58–69. 4

Phillips, T. J., Huson, M., Gwiazdon, C., Burkhart-Kasch, S., and Shen, E. H. (1995). Effects of acute and repeated ethanol exposures on the locomotor activity of BXD recombinant inbred mice. *Alcohol Clin Exp Res*, 19(2):269–78. 4, 54, 96

Pierucci-Lagha, A., Covault, J., Feinn, R., Nellissery, M., Hernandez-Avila, C., Oncken, C., Morrow, A. L., and Kranzler, H. R. (2005). GABRA2 alleles moderate the subjective effects of alcohol, which are attenuated by finasteride. *Neuropsychopharmacology*, 30(6):1193–203. 55

Pirrung, M. C., Fallon, L., and McGall, G. (1998). Proofing of photolithographic DNA synthesis with 3′,5′-dimethoxybenzoinyloxycarbonyl-protected deoxynucleoside phosphoramidites. *The Journal of Organic Chemistry*, 63(2):241–246. 23

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Stat Appl Genet Mol Biol*, 4:Article16. 28

Putman, A. H. (2008). *Genetic and genomic analysis of ethanol-induced anxiolysis*. PhD thesis, Virginia Commonwealth University, Richmond, Virginia. 19, 60, 94, 101, 120, 124

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–27. 25

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501. 28

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 20

Rajagopalan, D. and Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6):788–93. 11

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous snps: server and survey. *Nucleic Acids Res*, 30(17):3894–900. 119

Rask, E., Walker, B. R., Söderberg, S., Livingstone, D. E. W., Eliasson, M., Johnson, O., Andrew, R., and Olsson, T. (2002). Tissue-specific changes in peripheral cortisol metabolism in obese women: increased adipose 11beta-hydroxysteroid dehydrogenase type 1 activity. *J Clin Endocrinol Metab*, 87(7):3330–6. 127

Reich, T., Edenberg, H. J., Goate, A., Williams, J. T., Rice, J. P., Van Eerdewegh, P., Foroud, T., Hesselbrock, V., Schuckit, M. A., Bucholz, K., Porjesz, B., Li, T. K., Conneally, P. M., Nurnberger, Jr, J. I., Tischfield, J. A., Crowe, R. R., Cloninger, C. R., Wu, W., Shears, S., Carr, K., Crose, C., Willig, C., and Begleiter, H. (1998). Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet*, 81(3):207–15. 1

Reimers, M. (2010). Making informed choices about microarray data analysis. *PLoS Comput Biol*, 6(5):e1000786. 25

Reimers, M. A., Riley, B. P., Kalsi, G., Kertes, D. A., and Kendler, K. S. (2011). Pathway based analysis of genotypes in relation to alcohol dependence. *Pharmacogenomics J*. 11

Rimondini, R., Arlinde, C., Sommer, W., and Heilig, M. (2002). Long-lasting increase in voluntary ethanol consumption and transcriptional regulation in the rat brain after intermittent exposure to alcohol. *FASEB J*, 16(1):27–35. 10

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7. 2

Risinger, F. O. (2003). Genetic analyses of ethanol-induced hyperglycemia. *Alcohol Clin Exp Res*, 27(5):756–64. 89

Robinson, T. E. and Kolb, B. (1997). Persistent structural modifications in nucleus accumbens and prefrontal cortex neurons produced by previous experience with amphetamine. *J Neurosci*, 17(21):8491–7. 63

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8. 7, 11

Ruano, D., Abecasis, G. R., Glaser, B., Lips, E. S., Cornelisse, L. N., de Jong, A. P. H., Evans, D. M., Davey Smith, G., Timpson, N. J., Smit, A. B., Heutink, P., Verhage, M., and Posthuma, D. (2010). Functional gene group analysis reveals a role of synaptic heterotrimeric g proteins in cognitive ability. *Am J Hum Genet*, 86(2):113–25. 11

Saba, L., Bhave, S. V., Grahame, N., Bice, P., Lapadat, R., Belknap, J., Hoffman, P. L., and Tabakoff, B. (2006). Candidate genes and their regulatory elements: alcohol preference and tolerance. *Mamm Genome*, 17(6):669–88. 17

Saito, M., Szakall, I., Toth, R., Kovacs, K. M., Oros, M., Prasad, V. V. T. S., Blumenberg, M., and Vadasz, C. (2004). Mouse striatal transcriptome analysis: effects of oral self-administration of alcohol. *Alcohol*, 32(3):223–41. 10

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–23. 7

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 37(7):710–7. 126

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., Johnson, J. M., Rohl, C. A., van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S. C., Schuetz, E., Rushmore, T. H., and Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107. 18

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302. 14, 17

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70. 22, 23

Schuckit, M. A. (1984). Subjective responses to alcohol in sons of alcoholics and control subjects. *Arch Gen Psychiatry*, 41(9):879–84. 9

Schuckit, M. A. (1994). Low level of response to alcohol as a predictor of future alcoholism. *Am J Psychiatry*, 151(2):184–9. 9, 92

Scott, R. E., White-Grindley, E., Ruley, H. E., Chesler, E. J., and Williams, R. W. (2005). P2P-R expression is genetically coregulated with components of the translation machinery and with PUM2, a translational repressor that associates with the P2P-R mRNA. *J Cell Physiol*, 204(1):99–105. 12

Sen, S., Satagopan, J. M., Broman, K. W., and Churchill, G. A. (2007). R/qtldesign: inbred line cross experimental design. *Mamm Genome*, 18(2):87–93. 99

Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 6(7):639–45. 22

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504. 20

Shifman, S., Bell, J. T., Copley, R. R., Taylor, M. S., Williams, R. W., Mott, R., and Flint, J. (2006). A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol*, 4(12):e395. 6, 19, 69

Shirley, R. L., Walter, N. A. R., Reilly, M. T., Fehr, C., and Buck, K. J. (2004). Mpdz is a quantitative trait gene for drug withdrawal seizures. *Nat Neurosci*, 7(7):699–700. 6

Sieghart, W. (1994). Pharmacology of benzodiazepine receptors: an update. *J Psychiatry Neurosci*, 19(1):24–9. 95

Spanagel, R., Montkowski, A., Allingham, K., Stöhr, T., Shoaib, M., Holsboer, F., and Landgraf, R. (1995). Anxiety: a potential predictor of vulnerability to the initiation of ethanol self-administration in rats. *Psychopharmacology (Berl)*, 122(4):369–73. 93

Stillwell, E. E., Zhou, J., and Joshi, H. C. (2004). Human ninein is a centrosomal autoantigen recognized by crest patient sera and plays a regulatory role in microtubule nucleation. *Cell Cycle*, 3(7):923–30. 121

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55. 7, 12

Substance Abuse and Mental Health Services Administration (2011). *Results from the 2010 National Survey on Drug Use and Health: Summary of National Findings*, volume HHS (SMA) 11-4658 of *NSDUH Series H-41*. Department of Health and Human Services, Rockville, MD. 1

Taylor, B. A. (1978). Recombinant inbred strains: use in gene mapping. In Morse, H. C., editor, *Origins of inbred mice*, pages 423–438, New York. Academic Press. 5

Taylor, B. A., Wnek, C., Kotlus, B. S., Roemer, N., MacTaggart, T., and Phillips, S. J. (1999). Genotyping new BXD recombinant inbred mouse strains and comparison of bxd and consensus maps. *Mamm Genome*, 10(4):335–48. 5

Tolliver, B. K., Belknap, J. K., Woods, W. E., and Carney, J. M. (1994). Genetic analysis of sensitization and tolerance to cocaine. *J Pharmacol Exp Ther*, 270(3):1230–8. 88

Treadwell, J. A. and Singh, S. M. (2004). Microarray analysis of mouse brain gene expression following acute ethanol treatment. *Neurochem Res*, 29(2):357–69. 10

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21. 40

Urbanek, S. (2011). *multicore: Parallel processing of R code on machines with multiple cores or CPUs*. R package version 0.1-7. 158

Vadasz, C., Saito, M., Balla, A., Kiraly, I., Vadasz, 2nd, C., Gyetvai, B., Mikics, E., Pierson, D., Brown, D., and Nelson, J. C. (2000). Mapping of quantitative trait loci for ethanol preference in quasi-congenic strains. *Alcohol*, 20(2):161–71. 84

van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, 5(3):280–4. 61

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6. 7

Vengeliene, V., Bilbao, A., Molander, A., and Spanagel, R. (2008). Neuropharmacology of alcohol addiction. *Br J Pharmacol*, 154(2):299–315. 90

Wahlström, A., Hammar, L., Lundin, L. G., and Rane, A. (1986). Morphine metabolism in mouse brain. *NIDA Res Monogr*, 75:603–6. 88

Wallner, M., Hanchar, H. J., and Olsen, R. W. (2003). Ethanol enhances alpha 4 beta 3 delta and alpha 6 beta 3 delta gamma-aminobutyric acid type a receptors at low concentrations known to affect humans. *Proc Natl Acad Sci U S A*, 100(25):15218–23. 55

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2. 61

Wetherill, L., Schuckit, M. A., Hesselbrock, V., Xuei, X., Liang, T., Dick, D. M., Kramer, J., Nurnberger, Jr, J. I., Tischfield, J. A., Porjesz, B., Edenberg, H. J., and Foroud, T. (2008). Neuropeptide y receptor genes are associated with alcohol dependence, alcohol withdrawal phenotypes, and cocaine dependence. *Alcohol Clin Exp Res*, 32(12):2031–40. 3

Whitfield, J. B. (2002). Alcohol dehydrogenase and alcohol dependence: variation in genotype-associated risk between populations. *Am J Hum Genet*, 71(5):1247–50; author reply 1250–1. 3

Whitfield, J. B., Nightingale, B. N., Bucholz, K. K., Madden, P. A., Heath, A. C., and Martin, N. G. (1998). Adh genotypes and alcohol use and dependence in europeans. *Alcohol Clin Exp Res*, 22(7):1463–9. 3

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Use R! Springer, New York. 20

Williams, R. W., Bennett, B., Lu, L., Gu, J., DeFries, J. C., Carosone-Link, P. J., Rikke, B. A., Belknap, J. K., and Johnson, T. E. (2004). Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. *Mamm Genome*, 15(8):637–47. 5

Williams, R. W., Gu, J., Qi, S., and Lu, L. (2001). The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biol*, 2(11):RESEARCH0046. 19, 69

Wu, Z., Irizarry, R., and Gentleman, R. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909–17. 31

Yang, R. J., Mozhui, K., Karlsson, R.-M., Cameron, H. A., Williams, R. W., and Holmes, A. (2008). Variation in mouse basolateral amygdala volume is associated with differences in stress reactivity and fear learning. *Neuropsychopharmacology*, 33(11):2595–604. 88

York, T. P., Miles, M. F., Kendler, K. S., Jackson-Cook, C., Bowman, M. L., and Eaves, L. J. (2005). Epistatic and environmental control of genome-wide gene expression. *Twin Res Hum Genet*, 8(1):5–15. 14

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I. W., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Boström, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C.-J., Doney, A. S. F., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jørgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N. A., Payne, F., Perry, J. R. B., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Wellcome Trust Case Control Consortium, Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley, A. T., Collins, F. S., Groop, L.,

McCarthy, M. I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–45. 3

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4:Article17. 61, 124

Zhang, L., Miles, M. F., and Aldape, K. D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol*, 21(7):818–21. 30

Zhang, L., Wang, L., Ravindranathan, A., and Miles, M. F. (2002). A new algorithm for analysis of oligonucleotide arrays: application to expression profiling in mouse brain regions. *J Mol Biol*, 317(2):225–35. 31, 32, 33, 35, 40, 41

Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., Sachs, J. R., and Schadt, E. E. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol*, 3(4):e69. 48

Appendix A

R code

A.1 fishers_sscore

Function to identify ethanol responsive genes.

```
# fishers_sscore
################################################################################

# Identify Affymetrix probe-sets exhibiting significant expression variation
# by applying Fisher's combined probably test to S-score expression data.

# ARGUMENTS
# sscores: matrix of sscore expression data generated with the SScore package

# parallel: logical, If TRUE multicore package is loaded and all analysis
# is split among all available cores.

# n.core: number of cores to use for parallel execution. If unspecified all
# available cores will be used.

fishers_sscore <- function(sscores, n.perm, plot.results,
  parallel = FALSE, n.cores, verbose = FALSE){

  if(is.data.frame(sscores)) {
    sscores  <- as.matrix(sscores)
  }

  # Print status
  print_status <- function(message) {
```

```
  if(verbose) {
    dt <- format(Sys.time(), "%D %r")
    writeLines(paste("\n", dt, "-", message, "\n"))
  }
}

# Fisher's method to combine p-values
fishers_method <- function(data){
  S <- apply(data, M = 1, F = function(x) sum(-2 * log(x)))
  pvalue <- pchisq(S, df = ncol(data) * 2, lower.tail = F)
  return(data.frame(S, pvalue, row.names = rownames(data)))
}

# Probeset resampling function
permute <- function(x) {
  p <- sample(x, length(x), replace = F)
  return(p)
}

# Calculate empirical p-values from a matrix of permuted data
calc_emp <- function(obs, perm) {
  emp.p <- apply(perm, 2, function(x)
    sum(x > obs) / length(x))
  emp.p <- mean(emp.p)
  return(emp.p)
}

# Identify number of available cores and register parallel backend
if(parallel){
  print_status("Initializing multicore backend")
  require(multicore, quietly = T)
  require(doMC, quietly = T)
  available.cores <- multicore:::detectCores()

  if(available.cores < 2){
    parallel <- FALSE
    writeLines(paste("Sorry, you don't actually have a multicore CPU.\n",
      "Perhaps it's time to upgrade?\n\n", sep = ""))
  }

  if(missing(n.cores)){
    n.cores <- available.cores
  } else {
    if(n.cores > available.cores){
      n.cores <- available.cores
    }
  }
  registerDoMC(cores = n.cores)
}

# Convert sscores to pvalues
```

```
  print_status("Performing p-value transformation for observed data")
  pscores <- 2 * pnorm(abs(sscores), lower.tail = F)

  # Combine p-values
  print_status("Combining observed p-values")
  obs.results <- fishers_method(pscores)

  # Permutations
  ###############
  print_status(paste("Performing", n.perm, "permutations"))

  # Generate permuted matrix and calculate significance values
  # Return a list containing an entry for each permutation
  if(parallel) {
    perm.results <- foreach(p = 1:n.perm) %dopar% {
      perm.data <- apply(pscores, M = 2, F = permute)
      fishers_method(perm.data)
    }
  } else {
    perm.results <- list()
    for(p in 1:n.perm) {
      perm.data <- apply(pscores, M = 2, F = permute)
      perm.results[[p]] <- fishers_method(perm.data)
    }
  }

  # Calculate empirical pvalues
  ##############################
  print_status("Calculating empirical p-values")
  perm.s <- do.call("cbind", lapply(perm.results, function(x) x$S))

  if(parallel) {
    epval <- mclapply(obs.results$S, function(x) calc_emp(x, perm.s))
    epval <- unlist(epval)
  } else {
    epval <- sapply(obs.results$S, function(x) calc_emp(x, perm.s))
  }

  # Calculate observed and empirical qvalues
  obs.results$emp.pvalue <- epval
  obs.results$qvalue <- p.adjust(obs.results$pvalue, method = "fdr")
  obs.results$emp.qvalue <- p.adjust(obs.results$emp.pvalue, method = "fdr")

  print_status("Finished")
  return(obs.results)
}
```

A.2 ggAffy_ProbePlot

Function to visualize probe-level intensity data for multiple probe-sets from an AffyBatch object. Colors can be mapped to samples using `color.var` to specify a variable stored in the PhenoData slot.

```
# ggAffy_ProbePlot
################################################################################
# object: an AffyBatch-class
# probesets: character vector of probeset IDs.
# (warning: plot starts to become useless with > 5 probesets)
# color.var: variable to use for color mapping
# mm: logical, should mismatch probes be plotted
# fixed_yaxis: logical, should the y-axis be fixed across probe-sets
# log2: logical, should intensity be log transformed

# DETAILS
# color.var must be a character vector of length 1 that identifies a
# variable in the AffyBatch object's phenoData slot. For best results it should
# be a qualitative variable with fewer than 9 levels. An example use case is to
# color samples by genotype at a particular loci to look for probe x allele
# interactions.

# EXAMPLE
# data(Dilution)
# probesets <- sample(featureNames(Dilution), size = 3)
# ggProbePlot(Dilution, probesets, mm = T)
# ggProbePlot(Dilution, probesets, color.var = "scanner", mm = T)

ggAffy_ProbePlot <- function(object, probesets, color.var,
  mm = T, log2 = T, fixed_yaxis = F){

  require(affy)
  require(ggplot2)

  if(class(object) != "AffyBatch"){
    stop("object must be an AffyBatch object.", call. = F)
  }

  # Obtain specified probe-set's probe index and extract expression data
  pm.index <- indexProbes(object, "pm", genenames = probesets)
  exp <- melt(lapply(pm.index, function(x) intensity(object)[x,]))
  exp$type <- "pm"

  # Include mismatch data if specified
```

```
if(mm){
  mm.index <- indexProbes(object, "mm", genenames = probesets)
  mm.exp <- melt(lapply(mm.index, function(x) intensity(object)[x,]))
  mm.exp$type <- "mm"

  # Combine pm & mm data
  exp <- rbind(exp, mm.exp)
  exp$type <- factor(exp$type, levels = c("pm", "mm"))
}

# Rename columns and provide ordered probe labels
names(exp) <- c("index", "sample", "intensity", "probeset", "type")
exp <- ddply(exp, .(sample, probeset, type), transform,
  Probe = factor(1:length(index)))

# Maintain probesets order
exp$probeset <- factor(exp$probeset, levels = probesets)

# Add column for specified phenotypic variable
if(!missing(color.var)){
  if(!validObject(object@phenoData)){
    stop(paste("AffyBatch object must contain phenoData",
      "AnnotatedDataFrame in order to map colors to a variable." ), .call = F)
  }
  # Create new data frame aligning samples and color variable
  var.data <- data.frame(sample = sampleNames(object),
    variable = pData(object)[, color.var])
  # Add to exp data.frame
  exp <- merge(exp, var.data, by = "sample")
}

# log2 transform data
if(log2){
  exp$intensity <- log2(exp$intensity)
}

# Construct basic plot framework
exp.plot <- ggplot(exp) +
  aes(Probe, intensity, group = sample) +
  geom_line(alpha = .25) +
  opts(title = "Probe-level expression",
    plot.margin = unit(rep(0, 4), "lines"))

# Accommodate multiple probes if necessary
if(length(probesets) > 1){
  fixed_yaxis <- ifelse(fixed_yaxis, "fixed", "free_y")
  exp.plot <- exp.plot + facet_grid(probeset~., scales = fixed_yaxis)
} else {
  exp.plot <- exp.plot +
    opts(title = paste(probesets, "probe-level expression"))
}
```

```
  # yaxis label
  if(log2){
    exp.plot <- exp.plot + ylab(expression(log[2] * " intensity"))
  } else {
    exp.plot <- exp.plot + ylab("Raw intensity")
  }

  # Map linetype aesthetic to probe type
  if(mm){
    exp.plot <- exp.plot + aes(group = sample:type,
      linetype = type) +
      scale_linetype("Probe\ntype")
  }

  # Map color aesthetic to provided variable
  if(!missing(color.var)){
    if(nlevels(exp$variable) <= 9){
      exp.plot <-  exp.plot + aes(color = factor(variable)) +
        scale_color_manual(color.var, values =
          head(RColorBrewer::brewer.pal(9, "Set1"), nlevels(exp$variable)))
    } else {
      exp.plot <-  exp.plot + aes(color = factor(variable)) +
        scale_color_discrete(color.var)
    }
  }

  return(exp.plot)
}
```

A.3 ggAffy_Hist

Simultaneously visualizing the probe intensity histograms for all samples in a dataset is a very useful quality assessment procedure. Outlier samples can be easily spotted with this approach, even when their deviation from the dataset isn't so apparent with boxplots of probe intensities. As with the default `hist` function provided by the `affy` package, you can specify whether histograms should be generated using intensities for PM probes, MM probes or both. However, intensity histograms generated for large datasets typically suffer from over-plotting, making it difficult to determine which line corresponds to which sample. `ggAffy_Hist` provides a solution to this issue by allowing users to specify a subset of samples and comparing them to a reference array, which is added to the figure when `median.ref = TRUE`. The reference array is created by calculating probe-wise medians across the entire dataset, not just the subset group. This makes it possible to see how the distributions of each subset compare to the larger dataset. I typically use this function in a loop that generates one figure for every 8 samples; any more and the readability begins to suffer.

```
# ggAffy_Hist
###############################################################################
# object: an AffyBatch-class
# which: character, use pm probes, mm probes or both
# samples: optional character vector of samples indicating a subset of arrays
# in AffyBatch to plot
# log2: logical, should intensity be log transformed
# median.ref: logical, generate and plot a common psuedo-array reference

# DETAILS
# A vector of samples can be supplied to generate histograms
# for a subset of the dataset. Setting median.ref to TRUE will facilitate
# comparison of each of subset with the entire dataset by generating a
# synthetic array created by calculating probe-wise medians and plotting
# its intensity histogram.

# EXAMPLE
```

```
# data(Dilution)
# ggAffy_Hist(Dilution, which = "pm")
# ggAffy_Hist(Dilution, which = "both")

# sample.sub <- sampleNames(Dilution)[1:2]
# ggAffy_Hist(Dilution, which = "pm", samples = sample.sub, median.ref = T)

ggAffy_Hist <- function(object, which = "pm", samples, log2 = T, median.ref = F){

  require(affy, quietly = T)
  require(ggplot2, quietly = T)

  if(class(object) != "AffyBatch"){
    stop("object must be an AffyBatch object.", call. = F)
  }

  # Use all samples if no subset is provided
  if(missing(samples)){
    samples <- sampleNames(object)
  }

  # x-axis label
  x.lab <- paste(ifelse(which == "both", "PM/MM", toupper(which)),
    "probe intensity")

  # Extract raw probe intensities
  if(log2){
    exp <- log2(intensity(object))
    x.lab <- bquote(log[2] ~ .(x.lab))
  } else {
    exp <- intensity(object)
    x.lab <- paste("Raw", x.lab)
  }

  # Function to calculate kernel density for each column
  # and return a df with xy coordinates
  density_df <- function(exp){
    lst <- apply(exp, M = 2, F = density)
      d.x <- do.call(cbind, lapply(lst, function(x) x$x))
      d.y <- do.call(cbind, lapply(lst, function(x) x$y))
    return(data.frame(
      "x" = as.numeric(d.x),
      "y" = as.numeric(d.y),
      "sample" = rep(colnames(d.x), each = nrow(d.x))))
  }

  # Obtain probe indices and store in long format expression df
  probes <- unlist(indexProbes(object, which))
  dens_df <- density_df(exp[probes, samples])

  # Maintain original sample order
```

```
  dens_df$sample <- factor(dens_df$sample, levels = samples)

  # Generate a median chip to compare all arrays
  if(median.ref){
    mdn.chp <- density_df(data.frame(median = rowMedians(exp)))
    hist.plot <- ggplot() +
      geom_area(data = mdn.chp,
        aes(x = x, y = y, fill = sample), color=NA, alpha=.4) +
      scale_fill_manual("Dataset\nmedian", values = "grey50")
  } else {
    hist.plot <- ggplot()
  }

  # Render plot
  hist.plot <- hist.plot +
    geom_line(data = dens_df,
      aes(x = x, y = y, color = sample)) +
    xlab(x.lab) + ylab("Density")

  # Use Color Brewer Set1 palette if batch contains fewer than 9 samples
  if(length(samples) <= 9){
    hist.plot <- hist.plot +
      scale_color_manual("Sample", values =
        head(RColorBrewer::brewer.pal(9, "Set1"), length(samples)))

  } else {
    hist.plot <- hist.plot +
      scale_colour_discrete("Sample")
  }

  return(hist.plot)
}
```

A.4 snp_prober

As described in section 2.6.2, unaccounted for polymorphisms within microarray probe target regions may affect probe/target hybridization and skew reported measurements of transcript abundance. This collection of functions was implemented to make convenient the process of identifying problematic probes on an Affymetrix GeneChip microarray. These functions rely heavily upon data provided by Bioconductor and depend on the following packages:

- `BSgenome`
- `Biostrings`
- `GenomicRanges`
- `IRanges`

Optional arguments for `find_probeSNPs` include `probesets`, `num.mm`, `parallel` and `n.cores`. The `probesets` arguments makes it possible to specify a subset of probe-sets, by default all probe-sets that map to the specified chr are processed. Setting the `parallel` argument to `TRUE` will enable the most resource-intensive components of this process to be distributed across multiple central processing unit (CPU) cores through integration with the `foreach` package (Analytics, 2011). Doing so substantially reduces the amount of time required for this analysis to complete. By default, all available cores are used. As using all available cores is not always desirable and in some cases may affect system stability, the user may specify the number of cores to make available with the `n.cores` argument. The backend that `foreach` relies on is provided by the `multicore` package (Urbanek, 2011), which, as of this writing, only supports unix-like

operating systems like Mac OS X, Linux and Solaris. Therefore `snp_prober` cannot be run in parallel on Windows.

The `chr` argument is required because `snp_prober` currently supports matching to one chromosome at a time. Although I plan to remove this limitation in future versions, the matching process is so resource-intensive that a user will typically want to split up probe-sets into smaller batches anyway and doing so by chr is a natural strategy. The `probesets` argument allows the user to further subset the number of probes to process. Any probe-sets that do not map to the specified chr will be discarded.

```
# snp_prober
################################################################################

# Identify Affymetrix probe target sequences that are polymorphic relative
# to a list of user supplied alleles and positions.

# The position of each probe binding region is determined by aligning
# target sequence to genome.

# ARGUMENTS
# microarray: character, name of the affymetrix platform (eg mouse4302)
# probesets: optional character vector of probesets to query for SNPs,
# if provided any chr and chr.range information is ignored

# chr, chromosome that supplied probesets AND SNPs reside on
# this can handle only one chromosome at a time!

# snp.pos, numeric vector of SNP bp positions

# parallel, logical. If TRUE multicore package is loaded and all analysis
# is split among all available cores.

# n.cores, number of cores to use for parallel execution. If unspecified all
# available cores will be used.

# DETAILS
# The probe SNP analysis will be performed on a subset of probesets if a
# list of probesets is provided. Or on a subset of probesets that map to
# specificed chromosome and/or chromosome region. Otherwise the analysis is
# performed on all probesets available for affy platform.

# USER MUST SUPPLY VECTOR THAT SPECIFIES THE probeset chr location
```

```
# Each probe sequence is matched against the latest genome sequence data,
# and probe positions are returned from these perfect matches. Where
# a perfect match can not be found, probe positions are then deduced
# using the positions of their surrounding probes and labeled as inferred.

# num.mismatches;  every allowable mismatch causes the matching process to
  # take substantially longer

snp_prober <- function(microarray, chr, probesets, num.mm = 0,
  parallel = F, n.cores, install.missing = T, span.exons = F, attempt.bm = 3){

  # Required libraries
  require(BSgenome, quietly = T)

  if(missing(chr) | length(chr) > 1) {
    stop("You must specify a single chromosome.", call. = F)
  }

# Load bioconductor annotation packages for microarray
anno.pkgs <- load_annotations(
    microarray, type = c("db", "probe"), install.missing)
  db.pkg <- anno.pkgs["db"]
  probe.pkg <- anno.pkgs["probe"]

  # Determine organism and create BSgenome friendly name
  organism <- eval(as.name(paste(microarray, "ORGANISM", sep = "")))
  BS.organism <- gsub("(^\\w{1})\\w+\\s(\\w+)", "\\1\\2", organism)

  # Identify most recent genome package
  genomes <- available.genomes()
  genome.pkg <- sort(
    genomes[grep(BS.organism, genomes)], decreasing = T)[1]

  # Install genome package if neccessary
  if(!genome.pkg %in% installed.genomes()) {
    if(install.missing){
      writeLines(paste("\nInstalling ", organism, " genome package:\n",
        genome.pkg, "\n\nSit tight. This could take a while...\n", sep = ""))
      source("http://www.bioconductor.org/biocLite.R", verbose = F)
      biocLite(genome.pkg)
    } else {
      stop(paste("Please install the", organism, "genome package", genome.pkg,
        "\nor set install.missing = T to have it done for you."), call. = F)
    }
  }

  # Load genome package
  require(genome.pkg, character.only = T, quietly = T)

  # Ensure all neccessary packages are loaded
  all.pkgs <- as.character(c(db.pkg, probe.pkg, genome.pkg))
```

```
loaded <- all.pkgs %in% gsub("package:", "", search())

if(sum(!loaded)) {
  stop(paste("Missing the following packages:\n",
    paste(all.pkgs[loaded], collapse = "\n"), sep = ""), call. = F)
}

# Identify number of available cores and register parallel backend
if(parallel){
  require(multicore, quietly = T)
  require(doMC, quietly = T)
  available.cores <- multicore:::detectCores()

  if(available.cores < 2){
    parallel <- FALSE
    writeLines(paste("Sorry, you don't actually have a multicore CPU.\n",
      "Perhaps it's time to upgrade?\n\n", sep = ""))
  }

  if(missing(n.cores)){
    n.cores <- available.cores
  } else {
    if(n.cores > available.cores){
      n.cores <- available.cores
    }
  }
  registerDoMC(cores = n.cores)
}

# Create annotated data.frame for probe sequences
probe.data <- eval(as.name(probe.pkg))

probe.data <- probe.data[, -which(names(probe.data) == "Target.Strandedness")]

# Use xy coordinates as unique probe identifiers
probe.data$id <- with(probe.data, paste(x, y, sep = ""))

# Subset by probe-sets
if(!missing(probesets)) {
  probe.data <- subset(probe.data, Probe.Set.Name %in% probesets)
} else {
  probesets <- mappedkeys(eval(as.name(paste(microarray, "CHR", sep = ""))))
}

# Add chr info to probe.data
chr.map <- lapply(mget(probesets,
  eval(as.name(paste(microarray, "CHR", sep = "")))), function(x) x[1])
chr.map <- data.frame(Probe.Set.Name = names(chr.map), Chr = unlist(chr.map))
probe.data <- merge(probe.data, chr.map, by = "Probe.Set.Name")

# Subset probe.data to chromosome
```

```
probe.data <- subset(probe.data, Chr == chr)
probesets <- unique(probe.data$Probe.Set.Name)

# Get sequence data for chromosome
seq <- eval(as.name(BS.organism))[[paste("chr", chr, sep = "")]]

# Identify probe matches along genomic sequence
##############################################################################

# Use match_probe_seqs function for actual sequence matching
if(parallel){
  # Multicore analysis
  match.results <- foreach(p = probesets) %dopar% {
    cur.p <- subset(probe.data, Probe.Set.Name == p)
      data.frame(probeset = p, id = cur.p$id,
      match_probe_seqs(cur.p$sequence, seq, num.mm))
  }
} else {
  # Single core analysis
  match.results <- list()
  pb <- txtProgressBar(1, length(probesets), style = 3)
  for(p in probesets) {
    setTxtProgressBar(pb, which(probesets == p))
    cur.p <- subset(probe.data, Probe.Set.Name == p)
    match.results[[p]] <- data.frame(
      probeset = p, id = cur.p$id,
      match_probe_seqs(cur.p$sequence, seq, num.mm))
  }
}
match.results <- do.call("rbind", match.results)

# Merge probe.data with matches
match.results <- merge(
  match.results, probe.data[2:ncol(probe.data)], by = "id")

# Number probes based on position
match.results$probe.number <- unlist(with(match.results,
  tapply(Probe.Interrogation.Position, probeset, rank)))

# Reorder results
match.results <- match.results[with(match.results,
  order(probeset, probe.number)),]

match.results <- transform(match.results,
  strand = as.character(strand))

if(!span.exons){
  return(match.results)
}

# For unmatched probes, check for matching sequences that span exons
```

```
##############################################################################
unmatched <- subset(match.results, probe.start == 0 | probe.end == 0)

# Count matched probes per set
match.n <- aggregate(probe.start ~ probeset, match.results,
  function(x) sum(x > 0))

# Exclude probe-sets with zero matched probes
bad.sets <- as.character(subset(match.n, probe.start == 0)$probeset)
unmatched <- subset(unmatched, !probeset %in% bad.sets)

# Extract ensemblIDs from bioconductor annotation package
# (keep first id if multiple are returned)
unmatched$ensemblid <- unlist(lapply(mget(as.character(unmatched$probeset),
    eval(as.name(paste(microarray, "ENSEMBL", sep = "")))), function(x) x[1]))

# Exclude probe-sets without ensemblID
unmatched <- subset(unmatched, !is.na(ensemblid))

# Use ensemblIDs to download exon coordinates from biomart
# If no data is returned wait 60s and try attempt.bm more times
a <- attempt.bm

while(a > 0){
  exon.data <- get_ensembl_exons(unique(unmatched$ensemblid), organism)
  if(is.data.frame(exon.data)){
    a <- 0
  } else {
    warning(paste("\nUnable to connect to Biomart. Will make\n",
      "attempt ", (attempt.bm - a) + 1, " of ", attempt.bm,
      " in 60 seconds.\n", sep = ""), call. = F)
    a <- a - 1
    Sys.sleep(60)
  }
}

if(!is.data.frame(exon.data)){
  stop("No data could be retrieved from biomaRt.", call. = )
}

# Use search_spanning_exons  to identify cross-exon matches
e.ids <- unique(unmatched$ensemblid)

if(parallel){
  # Multicore analysis
  span.results <- foreach(e = e.ids) %dopar% {
    cur.e <- subset(exon.data, ensemblid == e) # current exon
    cur.p <- subset(unmatched, ensemblid == e) # current probeset

    matches <- with(cur.e, search_spanning_exons(
      start, end, seq, cur.e$strand[1], cur.p$sequence))
```

```
      if(!is.na(matches)){
        data.frame(probeset = cur.p$probeset[1], matches, id = cur.p$id)
      }
    }
  } else {
    # Single core analysis
    span.results <- list()
      for(e in e.ids) {
        cur.e <- subset(exon.data, ensemblid == e) # current exon
        cur.p <- subset(unmatched, ensemblid == e) # current probeset

        matches <- with(cur.e, search_spanning_exons(
          start, end, seq, cur.e$strand[1], cur.p$sequence))

        if(!is.na(matches)[1]) {
          span.results[[e]] <- data.frame(
            probeset = cur.p$probeset[1], id = cur.p$id, matches)
      }
    }
  }
  span.results <- do.call("rbind", span.results)

  # Add exon spanning matches to match.results
  if(nrow(span.results) > 0) {
    si <- which(match.results$id %in% span.results$id)

    match.results[si, "probe.start"] <- as.character(span.results$probe.start)
    match.results[si, "probe.end"] <- as.character(span.results$probe.end)
    match.results[si, "mismatches"] <- as.character(span.results$mismatches)
  }
  return(match.results)
}
```

A.4.1 load_annotations

Requires only the name of microarray platform to load all necessary annotation packages.

```
# load_annotations
################################################################################

# Loads bioconductor annotation data for specified microarray platform.

# ARGUMENTS
# microarray: character, unique identifier of microarray platform.
# type: character, annotation data type to retrieve.
# install.missing: logical, install package if missing.

# RETURNS
# Character vector of successfully loaded packages

load_annotations <- function(microarray, type = c("cdf", "db", "probe"),
  install.missing = F) {

  # Check microarray against list of annotation data available from Bioconductor
  source("http://www.bioconductor.org/biocLite.R", verbose = F)

  anno.url <- biocinstallRepos()[grep("annotation", biocinstallRepos())]

  annos <- available.packages(
    paste(anno.url, "src/contrib", sep = "/"))

  # Relevant packages for provided microarray
  annos <- rownames(annos)[grep(microarray, rownames(annos))]

  # Ensure microarray specifies a unique platform
  # and specified annotation types are available
  if(length(annos) > 0) {

    annos.type <- sapply(type, simplify = F, function(x) annos[grep(x, annos)])
    type.count <- unlist(lapply(annos.type, length))

    if(sum(type.count > 1)) {
      nonunique <- names(type.count)[type.count > 1]
      ambig.arrays <- gsub(nonunique[1], "", annos.type[[nonunique[1]]])
      stop(paste(microarray, " matches multiple platforms.\n\n",
        "Please use one of the following unique identifiers:\n",
        paste(ambig.arrays, collapse = "\n"), sep = ""), call. = F)
    }
    # Limit annos to include only those of specified data type
    annos  <- unlist(annos.type)
  } else {
```

```
    stop(paste("No annotation data found for", microarray), call. = F)
  }

  # Check if identified annotation packages are installed using
  # find.package, as installed.package can be quite slow
  installed <- sapply(annos, function(x)
    ifelse(length(find.package(x, quiet = T)), TRUE, FALSE))

  # Install missing packages
  if(sum(!installed)) {
    if(install.missing){
      biocLite(annos[!installed])

      # Recheck package status
      installed <- sapply(annos, function(x)
        ifelse(length(find.package(x, quiet = T)), TRUE, FALSE))
    } else {
      writeLines(paste("\nThe following packages could not be loaded because",
        " they are not installed on this system:\n",
        paste(annos[!installed], collapse = "\n"), sep = ""))
    }
  }

  # Load installed packages
  if(sum(installed)){
    loaded <- sapply(annos[installed], function(x)
      require(x, character.only = T, quietly = T))

    writeLines(paste("\nThe following packages were successfully loaded:\n",
          paste(annos[installed & loaded], collapse = "\n"), sep = ""))
    return(annos[installed & loaded])
  } else {
    stop("\nNo packages could be loaded", call. = F)
  }
}
```

A.4.2 match_probe_seqs

The primary pattern matching function of snp_prober. Probe sequences are matched against the reference genome provided by BSgenome and a data frame containing the coordinates of all matches within a Chr sequence are returned. Prior to the matching operations, the strand from which the probe-set target is transcribed is determined by comparing the number of successful matches on the positive-strand versus successes on the negative strand. The num.mm argument specified in snp_prober is passed to match_probe_seqs, and allows the user to indicate how many probe/reference mismatches are allowable. It should be noted that every additional allowable mismatch greatly increases the processing time required to identify matches. I generally set num.mm to 1.

```
# match_probes
##############################################################################

# Function requires probetable object and entire chromosome
# sequence data from BSGenome package. Probe table object is subset
# of probe.pkg for current probe-set.

# This is where the actual matching of probe sequence to genome
# sequence takes place.

match_probe_seqs <- function(seqs, ref.seq, num.mm = 0){

  require(Biostrings, quietly = T)

  # Convert vector of sequences into DNAStringSet
  probe.set <- DNAStringSet(seqs)

  # Place count and match pattern functions within lapply wrapper
  count_match <- function(x, ref.seq) {
    lapply(x, function(y) countPattern(y, ref.seq, max = num.mm))
  }

  find_match <- function(x, ref.seq) {
    lapply(x, function(y) matchPattern(y, ref.seq, max = num.mm))
  }
```

```
  # Look for matches across forward strand sequence
  f.matches <- count_match(probe.set, ref.seq)

  # If more than 2 probes have no matches, try the reverse complement
  if(sum(f.matches == 0) > 2){
    probe.set <- reverseComplement(probe.set)
    r.matches <- count_match(probe.set, ref.seq)
    # Gene is on negative strand if there are more rev matches
    if(sum(f.matches == 0) > sum(r.matches == 0)){
      strand <- "-"
    } else {
      # Stick with forward strand despite the failed matches
      strand <- "+"
      probe.set <- reverseComplement(probe.set)
    }
  } else {
    strand <- "+"
  }

  # Perform matching
  matches <- find_match(probe.set, ref.seq)

  # Replace integer(0) returned for failed matches with NA's so that
  # results data.frame is of proper size
  probe.start <- unlist(lapply(matches,
    function(x) ifelse(length(start(x)) == 0, 0, start(x))))

  probe.end <- unlist(lapply(matches,
    function(x) ifelse(length(start(x)) == 0, 0, end(x))))

  # Count mismatches
  mismatches <- sapply(1:length(probe.set), function(x)
    nmismatch(probe.set[[x]], matches[[x]]))

  mismatches <- unlist(lapply(mismatches,
    function(x) ifelse(length(x) == 0, NA, x)))

  return(data.frame(probe.start, probe.end, mismatches, strand))
}
```

A.4.3 get_ensembl_exons

Although the vast majority of probes target sequences within exons, a subset of probes target regions harboring exon-exon junctions. Because match_probe_seqs looks for probe sequence matches across genomic DNA, introns disrupt the alignment of exon spanning probes. This issue is handled by downloading the exon sequences for a probe target using get_ensembl_exons and repeating search across the assembled transcript sequence with search_spanning_exons.

```
get_ensembl_exons <- function(ensemblids, organism, strand = T) {

  # Load biomart
  require(biomaRt, quietly = T)

  # Alter organism to match biomart dataset format
  organism <- tolower(gsub("(^\\w{1})\\w+\\s(\\w+)", "\\1\\2", organism))

  # Connect to Ensembl BioMart database and organism-specific dataset
  db <- useMart("ensembl")
  mart <- useDataset(paste(organism, "gene_ensembl", sep = "_"), mart = db)

  # Retrieve exon locations and sequences
  exons <- getBM(attributes = c("ensembl_gene_id",
    "gene_exon", "exon_chrom_start", "exon_chrom_end"),
    filters="ensembl_gene_id", values = ensemblids, mart = mart)

  # Return NA if no data was retrieved from biomart
  if(identical(nrow(exons), 0L)){
    return(NA)
  }

  names(exons) <- c("seq", "ensemblid", "start", "end")

  # Order exons
  exons <- exons[with(exons, order(ensemblid, start)),]

  # Identify transcript strand
  if(strand){
    strand <- getBM(attributes = list("ensembl_gene_id", "strand"),
      filter = "ensembl_gene_id",
      values = unique(exons$ensemblid), mart = mart)
    names(strand) <- c("ensemblid", "strand")
```

```
    # Add strand to exons
    exons <- merge(exons, strand, by = "ensemblid")
    exons <- transform(exons, strand = ifelse(strand == 1, "+", "-"))
  }
  return(exons)
}
```

A.4.4 search_spanning_exons

```
search_spanning_exons <- function(starts, ends, seqs, strand,
  probe.seqs, num.mm = 0) {

  require(Biostrings, quietly = T)

  # Construct exon data.frame
  exons <- data.frame(
    start = starts, end = ends, seq = seqs, stringsAsFactors = F)

  # Convert vector of sequences into DNAStringSet
  probe.set <- DNAStringSet(probe.seqs)

  # If transcribed from negative strand convert sequences to reverse complement
  # so start/end positions will correctly correspond positive strand sequence
  if(strand == "-"){
    exons$seq <- sapply(exons$seq, USE.NAMES = F, function(x)
      as.character(reverseComplement(DNAString(x))))
    probe.set <- reverseComplement(probe.set)
  }

  # In the presence of multiple transcripts per gene, overlapping or redundant
  # exons are frequently returned and must be consolidated.
  # Check for exon overlap by creating a vector of all exon positions and
  # looking for recurring positions
  coords <- apply(exons, M = 1, function(x) x["start"]:x["end"])

  # Generate non-redundant exon coordinates and sequences
  if(max(table(unlist(coords))) > 1) {
    exons <- with(exons, consolidate_exons2(start, end, seq))
  }

  # Number exons and add column for sequence length
  exons <- transform(exons, number = order(start), length = nchar(seq))

  # Concatenate exons sequences into a single transcript
  transcript <- paste(exons$seq, collapse = "")

  # Transcript data.frame where each row corresponds to a single nucleotide
```

```
transcript.data <- data.frame(
  exon = as.numeric(unlist(apply(exons, M = 1, function(x)
    rep(x["number"], x["length"])))),
  chr.pos = as.numeric(unlist(apply(exons, M = 1, function(x)
    x["start"]:x["end"]))))

# Add within exon nucleotide positions
transcript.data$exon.pos <- unlist(with(transcript.data,
  tapply(chr.pos, exon, rank)))

# Match each probe sequence to the transcript
matches <- lapply(probe.set, function(x)
  matchPattern(x, DNAString(transcript), max = num.mm))

# Identify start and stop positions of each match relative to the transcript
# (Replace integer(0) returned for failed matches with NA's so that
# results data.frame is of proper size)

match.pos <- data.frame(match = 1:length(matches),
  start = unlist(lapply(matches,
    function(x) ifelse(length(start(x)) == 0, NA, start(x)))),
  end = unlist(lapply(matches,
    function(x) ifelse(length(start(x)) == 0, NA, end(x)))))

# Return NA coordinates if no matches found
if(sum(!is.na(match.pos$start)) == 0) {
  return(NA)
}

# Identify the exons spanned by the matched sequence so the nt
# positions that denote the begining and end of the match can be determined
match.pos <- transform(match.pos,
  exon1 = transcript.data$exon[start], exon2 = transcript.data$exon[end],
  exon1.start = transcript.data$chr.pos[start],
  exon2.end = transcript.data$chr.pos[end])

# Chromosomal coordinates of matches within the first and second exons
exon1 <- paste(match.pos$exon1.start,
  aggregate(chr.pos ~ exon,
  subset(transcript.data, exon %in% match.pos$exon1), max)$chr.pos, sep = "-")

exon2 <- paste(aggregate(chr.pos ~ exon,
  subset(transcript.data, exon %in% match.pos$exon2), min)$chr.pos,
  match.pos$exon2.end, sep = "-")

exon1[grep("NA", exon1)] <- NA
exon2[grep("NA", exon2)] <- NA

# Count mismatches
mismatches <- sapply(1:length(probe.set), function(x)
  nmismatch(probe.set[[x]], matches[[x]]))
```

```
  mismatches <- unlist(lapply(mismatches,
    function(x) ifelse(length(x) == 0, NA, x)))

  return(data.frame(probe.start = exon1, probe.end = exon2, mismatches))
}
```

A.4.5 consolidate_exons

The exon sequence data obtained from Ensembl is often massively redundant due to the existence of multiple isoforms or splice variants. consolidate_exons does what it says and returns a single, representative transcript that encompasses all other exons. Figure A.1 provides a visualization of a consolidated group of exons.

```
consolidate_exons2 <- function(starts, ends, sequences, plot = F, colors){

  # Build data.frame with original exon positions
  orig.exons <- data.frame(start = starts, end = ends)

  # Identify chromosomal coordinates for consolidated exons
  ##############################################################################

  # Split exons into indivdiual data.frames, where each row = 1 base
  if(missing(sequences)){
    pos.mat <- apply(orig.exons, 1, function(x)
      data.frame(pos = x["start"]:x["end"]))
  } else {
    orig.exons$seq <- sequences
    pos.mat <- apply(orig.exons, 1, function(x)
      data.frame(pos = x["start"]:x["end"],
      seq = strsplit(x["seq"], split = ""),
      stringsAsFactors = F))
  }

  # Compile vector of unique exon positions
  exon.pos <- sort(unique(unlist(lapply(pos.mat, function(x) x$pos))))

  # Complete positions sequence from begining of first exon to end of last exon
  all.pos <- seq(min(exon.pos), max(exon.pos))

  # Denote intron positions by N
  all.pos <- paste(replace(
    all.pos, !all.pos %in% exon.pos, "N"), collapse = "-")

  # Compress each intron to a single N
  all.pos <- gsub("(-N)+", "-N", all.pos)

  # Consolidated exon start positions
  new.starts <- gregexpr("^\\d+|N-\\d+", all.pos)[[1]]
  new.starts <- substring(all.pos, new.starts,
    new.starts + attr(new.starts, "match.length"))
  new.starts <- as.numeric(gsub("\\D*", "", new.starts))
```

```
# Exon end positions
new.ends <- gregexpr("\\d+-N|\\d+$", all.pos)[[1]]
new.ends <- substring(all.pos, new.ends,
  new.ends + attr(new.ends, "match.length"))
new.ends <- as.numeric(gsub("\\D*", "", new.ends))

# New exon data.frame
new.exons <- data.frame(start = new.starts, end = new.ends)

# Construct sequences for consolidated exons
#############################################################################

# Iteratively merge each exon sequence into a single, consolidated transcript
if(!missing(sequences)){
  for(m in 1:length(pos.mat)){
    if(m == 1){
      cons.mat <- pos.mat[[m]]
    } else {
      cons.mat <- merge(cons.mat, pos.mat[[m]], all = T, by = c("pos", "seq"))
    }
  }
  # Add each consolidated exon's proper stretch of sequence
  new.exons$seq <- apply(new.exons, 1, function(x) paste(with(cons.mat,
    seq[pos >= x["start"] & pos <= x["end"]]), collapse = ""))
}

# Plot positions of consolidated exons against original exons
#############################################################################
if(plot){
  # Calculate minor gridlines
  calc_minor <- function(pos){
    int <- (pos[2] - pos[1]) / 2
    return(seq(min(pos) - int, max(pos) + int, int * 2))
  }

  if(missing(colors)){
    colors <- c("yellow", "#3B4FB8")
  }

  # Number exons
  orig.exons <- transform(orig.exons, number = order(start))

  # Base plot
  with(orig.exons, plot(NA, xlim = range(pretty(c(start,end))),
    ylim = range(number), axes = F,
    xlab = "Position (Mb)", ylab = "Exon"))

  # Background
  usr <- par("usr")
  rect(usr[1], usr[3], usr[2], usr[4], col = "grey90", border = NA)
```

```
    # Grid
    abline(v = calc_minor(axTicks(1)), h = calc_minor(axTicks(2)),
      col = "grey95", lwd = .75)
    abline(v = axTicks(1), h = axTicks(2), col = "white")

    # Axes
    axis(1, at = axTicks(1), labels = axTicks(1)*10^-6, lwd = 0, lwd.tick = 1)
    axis(2, at = axTicks(2), lwd = 0, lwd.tick = 1, las = 2)

    # Consolidated exons
    with(new.exons,
      rect(start, usr[3], end, usr[4], col = colors[2], border = NA))

    # Original exons
    with(orig.exons,
      rect(start, number + .3, end, number - .3, col = colors[1], border = NA))

    legend("bottom", bty = "n",
      legend = c("Original exon", "Consolidated exon"),
      col = colors, pch = 15, pt.cex = 1.5, cex = .85,
      horiz = T, xpd = T, inset = 1, xjust = 1)

  }
  return(new.exons)
}
```

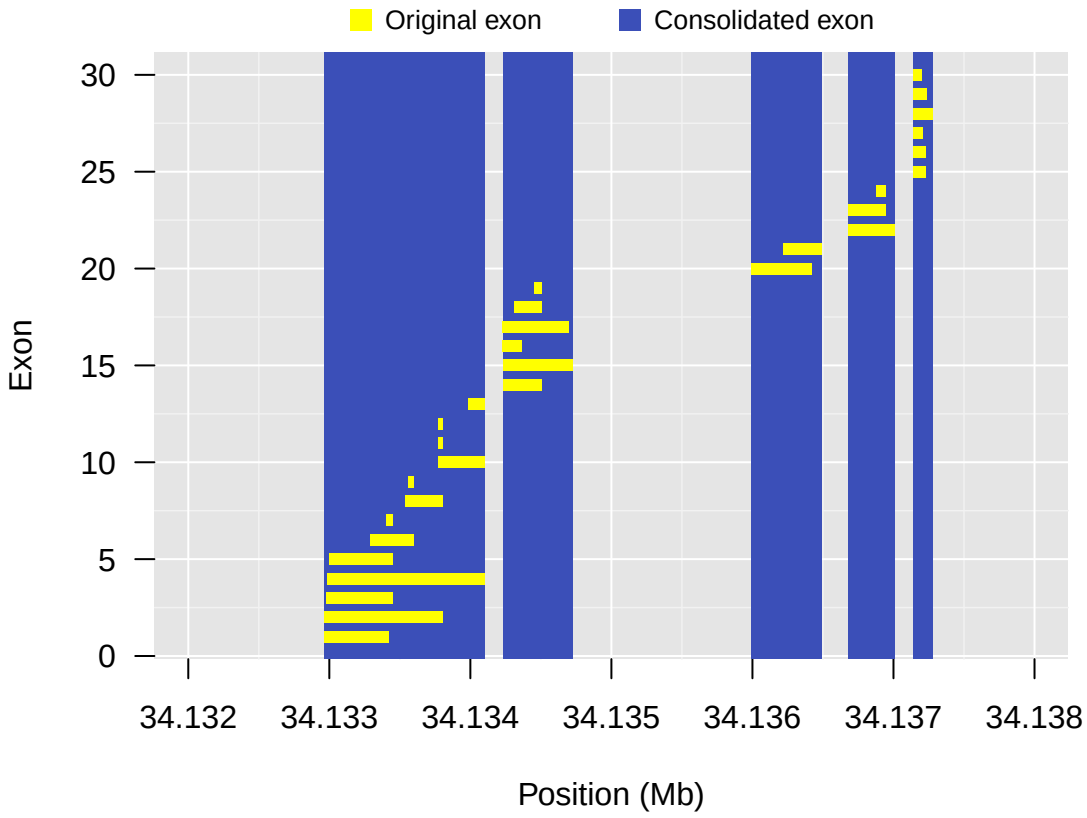

Figure A.1. A typical result produced by the `consolidate_exons` function. Yellow blocks represent the original exon sequences obtained from Ensembl, while purple blocks represent the consolidate product.

Appendix B

Supplemental tables

All supplemental data can be downloaded at aaronwolen.com/thesis, including the following supplementary tables.

B.1 Table S1

Lists of genes found to be significantly ethanol responsive in PFC, NAc and VMB by the analysis described in the Ethanol responsive genes across BXD panel section.

B.2 Table S2

Full results from functional over-representation analysis of ethanol responsive genes in PFC, NAc and VMB, discussed in section 2.4.4.

B.3 Table S3

Lists of genes found to be significantly ethanol responsive in PFC across the LXS panel as part of the analysis in section 2.5.

B.4 Table S4

List of genes that belong to the paraclique networks defined in the Saline versus ethanol S-score paraclique networks section using saline versus ethanol S-scores, as well as the saline and ethanol RMA data for the PFC data-set. Degree of connectivity and betweenness centrality measures are provided for each probe-set.

B.5 Table S5

Overlap of paraclique networks constructed using S-score data and the saline/ethanol RMA data-sets, as described in section 3.2.3.

B.6 Table S6

Peak eQTL results for all members of the saline RMA and S-score paracliques (section 3.2.3) with at least one suggestive eQTL found in the analysis discussed in section 3.3.3.

B.7 Table S7

Ranking results of positional candidate genes within each of the major ErGeN *trans*-bands.

B.8 Table S8

Full results from functional over-representation analysis of S-score networks identified in the PFC.

B.9 Table S9

List of Affymetrix M430v2 probe-sets that overlap one or more B6D2 polymorphisms.