Genomics of Complex Traits: Methods and Applications

by

Shweta Ramdas

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2018

Doctoral Committee:

Professor Jun Li, Co-Chair
Professor Margit Burmeister, Co-Chair
Assistant Professor Johann Gagnon-Bartsch
Associate Professor Hyun Min Kang
Associate Professor Cristen Willer

Shweta Ramdas

sramdas@umich.edu

ORCHID iD: 0000-0001-8888-4661

To my parents and teachers

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors in graduate school. Professor Jun Li has been a mentor who has constantly inspired me with his attitude to science; I always walk away from meetings with him motivated about research, and I'm very thankful for the time he has given me. I hope I become half the scientist he is. Dr Margit Burmeister has been a mentor since I joined Michigan, and her broad perspective on research and her insights on various topics in psychiatric genetics are always things I look forward to. She has also been extremely generous in pointing me towards opportunities of interest. Both Margit and Jun have shown me the utmost kindness when I have struggled with my mental health, and I cannot thank them enough for this. When I grow up, I want to be like Jun and Margit!

I have had the privilege of working with a wonderful thesis committee–Dr Cristen Willer, Dr Hyun Ming Kang and Dr Johann Gagnon-Bartsch, besides being scientists whose work I deeply admire, have always given me good feedback, and have been gentle with their critiques of my work. Dr Willer, especially, has been very generous with her time and advice every time I have sought her out.

I've also had the opportunity to learn from various collaborators and mentors in grad school—Dr Ryan Welch, Dr Tanya Teslovich, Dr Laura Scott, Dr Charles Burant, Dr Jessica Schwartz, Dr Megan Hagenaeur. I am grateful to them for their time and mentorship.

The Li lab has been very supportive with feedback and during setbacks. Big thanks to Ayse Bilge Ozel, Qianyi Ma, Wang Yu, Yu-yu Ren, Bo Li, Val Schaibley,

Brittany Nelson, Qingxuan Song, Xianing Zheng, Sahar Ansari, Simone Marini, Qing Ma, Jonhn Lloyd. And thank you for letting me eat all your food!

The wonderful administrative staff who have been pillars of support to all us grad students—in particular, Julia Eussen, Jane Weisner, Susan Kellogg, Dawn Keane, Alex Terzian.

Wonderful colleagues and journey(wo)men—Shiya Song, Sai Chen, Bryan Moyers, Xuefang Zhao, Wei Zhou, Brittany Nelson, Ray Cavalcante, Ellen Schmidt, Laura Seaman—this wouldn't have been nearly as much fun without their companionship. Faculty in DCMB and Human Genetics have always been generous with their time and advice, both scientific and personal. Nisha Patel and Deeksha Pai, whom I had the opportunity to mentor,have been wonderful students who have made me love what I do even more.

To my family, for their constant goodwill and blessings. Special thanks to Jaishna and Eisha Khullar, for being the ray of sunshine in each day.

Dr Scott Scholz and Vroni Sachsenhauser, whose calm, balance and wisdom have taught me so much everyday, and who make every day better by their presence. It has been a privilege to share my years in grad school with two people for whom I have unending admiration and love.

Ray Cavalcante and Teal Guidici, to whom I owe a big debt of gratitude for being there, always, especially when I didn't know I needed it. For nourishment that's intellectual, emotional, and *material* (thanks for all the dinners!) Megh Marathe, for making me laugh till I cry. Prachi Jalan, Anupama Yadav, for the sisterhood and for the constant vote of confidence. Akshat Agarwal, Amrita Srivathsan, Priyam Trivedi, Bharath Narayanan, Sandra Siby, Aswin Saravanan, Parul Tomar, Debjyoti Bardhan, Rahul Dandekar, Sharanya Murali, Dr Siddhika Iyer, Shiya Song, Dr Rachel Wilcox—thank you always for helping me pick up the pieces. Rushi Padhuman, Arun Chavan, Archana Reddy, Vaishali Sivakumar, Aashraya Ramu, Anshika Srivastava, Bilge Ozel, Ramya Kumar, Suvi Gunda, Amanda Pendleton—thank you for the friendship.

Mama, for the spiritual guidance and for being my moral compass, even if I fail to align with it properly. Amma and Appa, in whose shadows I will always walk.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

My dissertation covers a number of studies that seek to understand the genetics and functional basis of complex traits in either human or animal models. The first project is a genetic study of bipolar disorder using exome sequencing, primarily involving variant burden analysis and pathway analysis. The second project uses gene expression as a functional readout of a tissue system, which allows us to identify aging signatures in the eye. These two studies represent complementary approaches; while the first reflects inheritance patterns of DNA variants and phenotypes, and the second is a functional readout of an organ system. In the next two studies, I use a rat model involving both genetics and genomic tools to perform an integrative search for genes and functional pathways implicated in metabolic phenotypes. In this collaborative study, I combine multiple datasets including genotype-based QTL mapping and gene expression based functional comparison, seeking to triangulate signals that may be noisy or subtle in one platform alone. Along the way, I had to develop methodologies for data integration, and worked on solidifying existing resourcesin this case, the rat reference genome. Mine is the beginning of a consolidated effort by the rat-genomics community to arrive at a more complete and accurate reference genome, so that the community can build on this improved resource for more accurate research in the future. As my studies have involved both genome and exome sequencing data, one of the challenges is to identify not only single nucleotide variants but also larger scale DNA copy number changes. Despite the many tools for calling copy number changes,

there is still confusion about the proper investment of resources based on a principled power analysis. In my last chapter, I develop a basic mathematical framework that incorporates most of the important practical parameters impacting power, and create an online calculator and examples of some usage cases. Taken together, this dissertation is a reflection of how the field of genetics and genomics has moved in the last few years, involving rapidly advancing technology, and datasets with complex structures, requiring careful exploration and method development.

CHAPTER I

Introduction

1.1 History of Complex Trait Genetics

In the early twentieth century, researchers in the famed ‘fly room’ at Columbia University were performing the world’s first ever gene-mapping experiments on fruit flies. They performed crosses (‘arranged marriages’) between flies with known combinations of observable traits (or phenotypes); they then observed the relative frequencies of these traits in their offspring, and mapped relative positions of genes on chromosomes. They identified the first ever causal ‘genes’—fragments on chromosomes that they could map to a trait, finding the first mechanisms of inheritance. In humans, the first studies on variation came in 1919, when Hirschfield and Hirschfield [62] used variation in response to antibodies as surrogates for genetic variation. And these were before we knew the physical or chemical basis of heredity!

Once DNA was identified as the physical unit of heredity, and Franklin, Watson, Crick, and Wilkins elucidated its physical structure in 1953, a chain of fundamental discoveries helped create the field of molecular biology. The field of genetics largely merged with this field in the following decades. ‘Gene mapping’ for disease traits became popular amongst human geneticists. The first such studies used linkage analyses to narrow in on chromosomal locations for causal genes. Linkage analysis studied the segregation of genetic markers with disease status in families. Linkage analyses had poor resolution because of limited number of meiosis in families, and low density of the first generation of genetic markers, and was thus followed by an approach called positional cloning for further narrowing in on the gene. This approach was remarkably successful in identifying causal genes for many monogenic diseases. However, most such studies applied to complex diseases were undercut by the problems of incomplete penetrance and variable expressivity—where a mutation in a gene was neither necessary nor sufficient to cause the disease.

The Mendelian view of genetics was prevalent for much of the 19th century and early 20th century; most traits were expected to have a single causal gene, or in the case of added complexity, a few causal genes (numbers that today, seem quaint, as described by [17]. However, Mendel's phenotypes in peas were the exceptions; most biological traits—like height, skin color, and educational attainment—show a range of values. In 1918, Ronald Fisher helped resolve the conflict between the particulate nature of inheritance and the continuous distribution of most studied phenotypes. He referred to biological traits as pointillist paintings, with many factors (multiple genes plus the environment) coming together to contribute to the final phenotypic value [100].

In the 1950s, Victor McCusick at Johns Hopkins University laid down four important principles about clinical genetics that formalized some important principles of complex trait etiology (which is ironic given he is best known for his database of *Mendelian* genes in man.) First, mutations in a single gene can cause different manifestations of disease in different organs. Second, a single aspect of physiology can be influenced by multiple genes. Third, mutations in genes can have incomplete penetrance and variable expressivity. Fourth, mutations are just variations, and have no inherent 'value' or hierarchy of goodness. The first three would be of great value to the geneticist community in their hunt for complex trait genes, and the fourth for geneticists in relation to society at large.

The limitations of linkage analysis to the study of most traits eventually became apparent—effect sizes of individual causal variants were too small to detect via co-segregation within pedigrees [138]. As a solution, scientists used candidate-gene association studies, which looked for differential frequencies of genetic variants in genes hypothesized to be important for the studied phenotype. These studies were almost entirely un-replicable, and threw up a deluge of results that turned out to be false positives. In 1996, Risch and Merikangas quantitatively showed that association studies with more genetic markers and greater sample sizes would be more powerful than traditional family-based linkage analyses to detect common variants of small effect [120]. This and other papers [116] lay the groundwork for a new experimental paradigm in genetics—the genome-wide association study—which would become possible a little less than 10 years after it was originally proposed.

Genomic technology has revolutionized the study of complex traits. Different technologies have allowed us to ask new biological questions, and use new approaches to study existing ones—I elucidate these in section 1.2. The first genomic studies used restriction fragment length polymorphisms (RFLP), multi-nucleotide markers that are distantly placed signposts on the genome [16]. After the sequence of the human genome was released, the community moved on to genotyping arrays—chips that allow us to probe genetic variants common in the human population. Currently, the field is making the shift towards high-throughput sequencing technologies. The field of genetics has moved from the era of hypothesis-driven candidate gene approaches where a single gene would be tested in families for linkage to a disease, to studies on a genome-wide scale, looking across the genome or the exome to narrow in on potentially causal genes for further study ('hypothesis-generating').

We now see large population cohorts aimed at identifying causal genetic variants of small/large effect, family-studies aimed at identifying rare variants of larger effects. Our view of "genes as destiny" has now been conclusively repudiated for most traits, and we are seeing that different affected individuals for the same trait are likely to carry a unique portfolio of genetic risk factors, which manifest in the phenotype in the presence of certain environmental conditions. Most of these risk factors are in the non-coding region of the genome, which has been a stumbling block to interpretation, and has made essential the integration of DNA sequence data with other dimensions of biological information, such as gene expression and epigenetic information (section 1.3, 1.4). We are increasingly sequencing hundreds of thousands of people, but are yet to solve the problem of 'missing heritability'.

Complex trait genetics remains a field with many unsolved questions, particularly about the genetic architecture underlying different traits. I have elucidated some of the emerging principles in section 1.5. In this thesis, I have attempted to untangle regions of this hairball from various angles. While I describe each of my approaches in detail in each chapter, I briefly give an overview in the sections below. In Chapter 2, I study families falling along the near-Mendelian spectrum of complex disease using exome sequencing data, described in section 1.1. In Chapter 3, I study gene expression (Section 1.2) as a biomarker of a complex trait such as aging, and in Chapter 4, I study a rat model of aerobic capacity in the hopes of reducing some of the environmental complexities that are brought on when studying human populations. In Chapter 5, I analyze the rat genome assembly, which is a fundamental resource upon which any genomic analyses using a rat model depend, and discover technical flaws in the assembly that require addressing by the rat genome community. In Chapter 6, I build a tool to enable researchers to optimally design their experiments to have the maximum power to detect copy-number variations from high-throughput sequencing data from single cancer samples.

1.2 Genetic Data

In 2001, the human genome was sequenced. Since then, we have reference genomes for multiple organisms; these genomic maps have become essential tools in the geneticist's toolbox. Their quality and validity is taken for granted by most biologists; they are used as scaffolds upon which sequences from samples are aligned. I show in Chapter V that the reference genome of the rat has previously-overlooked defects, which need to be addressed by the community of rat-geneticists (geneticists studying rats, not rats who happen to be geneticists).

The first generation of truly genome-wide data came with the advent of SNP (Single Nucleotide Polymorphism) arrays. These are collections of DNA probes attached to a solid surface. These probes typically capture 500,000 to 1 million positions in the genome that represent "common variants". Linkage disequilibrium between nearby positions in the genome ensures that these common variants capture 80% of the variation in the genome. GWAS represent an implementation of the "common disease common variant" hypothesis, which states that common disease-causing variants will be found. Such chips have been used to carry out genome-wide association studies (or GWAS) in human studies, in which we look for SNPs that occur more commonly in cases than controls. One of the first GWAS was on acute macular degeneration (AMD) in 2005 was run on a sample of 96 cases and 50 control samples, and identified 1 locus in the *Cfh* gene associated with the disease [73]; a GWAS in 2016 for the same phenotype was run on 16,144 people and revealed 52 independent loci associated with the disease [43]. The intervening years have seen more than 3,000 GWAS on a plethora of phenotypes.

The main picture that has emerged from GWAS is our understanding of complex diseases was grossly oversimplified. Instead of finding loci of large effect on the phenotype, GWAS have revealed 100s of loci implicated in each phenotype, each of which contribute to a minute fraction of the heritability (analog causes on the phenotype instead of digital). As Nisbett et al say about intelligence, the number of genes involved in an outcomeis very large, and therefore the contribution of any individual locus is just as small as the number of genes is large [105]. The initial years of GWAS highlighted the importance of large sample sizes required to detect variants of small effect.

GWAS is often seen as a failure of the genetics community by many corners—for its failure to explain more of the phenotypic variance, and for being largely unable to deliver meaningful, biologically relevant knowledge or results of utility [138]. It failed to live up to its promise of delivering clinically actionable results, since common variants tended to increase the odds of disease by very small numbers, too small to be clinically meaningful. However, the ways GWAS have 'failed' have been illuminating. GWAS results point to a highly polygenic model of disease susceptibility with causal variants across the entire frequency of the allele frequency spectrum. They have also made clear that pleiotropy is commonmany variants are associated with multiple traits. Moreover, since 90% of significant GWAS loci fall in non-coding locations of the genome, it has become clear that changes in gene regulation, rather than changes to proteins, underlie most GWAS associations. There has been a concerted effort by the scientific community to shine a light on non-coding regions of the genome. Finally, SNP-based estimates of heritability show much lower values than heritability estimates from family-based studies [40], which tell us that we may have grossly over-estimated the genetic contribution to the variance of traits (while under-estimating environmental contributions). It is the 'failure' of GWAS that has uncovered a question that we didn't know existed—that of the missing heritability [122].

Although GWAS are unbiased with respect to prior biological knowledge, they are not unbiased in terms of what is detectable (common variants) [138]. A full appreciation of the genetic architecture of common disease requires an understanding of the role of rare variants. This has motivated a wave of rare variant association studies, which include whole-genome sequencing as well as whole-exome sequencing [8]. Rare variants are predicted to have larger effects than higher frequency SNPs, because we hypothesize greater selection acting against these variants, thus accounting for their low frequency. Despite accounting for a smaller percentage of the heritability (by being rarer), may allow us to pinpoint causality. The study of rare variants, however, brings with it a new set of statistical and computational challenges, since individual variants are too rare to have the statistical power to be tested. The standard approach in such studies is to aggregate rare variants into genes or pathways, or other biologically relevant units, and then use those as the unit of association. These studies have added to the polygenic signal from GWAS. Moreover, trio-based designs have allowed us to find new patterns of inheritance, as in autism, where we see an increased burden of de novo mutations in affected offspring. The number of sequenced samples promises to explode over the next decade with samples from public BioBanks [20], and reveal a comprehensive catalog of rare variants in the human population.

In parallel, sequencing studies have given rise to a rebirth of family-based studies, which are a complement to population approaches. We can now study segregation of rare variants in affected pedigrees, alleviating problems of population stratification, and environmental and phenotypic heterogeneity that often crop up in case-control studies. I describe our own efforts in identifying causal variations for a complex trait—bipolar disorder—using family-based sequencing approaches in Chapter 2.

1.2.1 Experimental Design in Sequencing Studies

Sequencing cost is still a bottleneck, so researchers must often make a trade-off between read depth, number of samples sequenced, and region of the genome to cover. Sequencing studies have been slow to adopt power analyses to experimental design. An analytical framework allows us to optimize experiments depending on the biological hypothesis we're testing, given a fixed cost. In Chapter 6, I develop a framework for optimizing experimental parameters to detect copy number variations (CNVs) from sequencing data.

1.3 Gene Expression, or Transcriptomics

While genetic variation is the blueprint of biology, we ultimately want to know how genes function in cells, and in an intact organism. Gene expression, measured by 'counting' some measure of the mRNA content in a cell, is a measure of gene function that has become popular in genomics. Gene expression is known to be a heritable phenotype from model organisms [27]. This measure then allows us to find intermediate links between genetic variation and phenotype. It is in no way the perfect metric of gene function (studies have shown low correlation between gene expression and protein levels [52], and mRNA levels do not reflect variation in post-translational modifications or protein localization), but the lack of high-throughput proteomic or metabolic approaches has made gene expression the readout of choice.

The first genome-wide studies of gene expression used chips (microarrays) with oligonucleotide probes specific to segments of mRNA. RNA from cells of interest was extracted, converted to cDNA, labeled, and then washed over the chip where the labeled cDNA molecules would bind to the probes. After unbound cDNA were washed away, the fluorescent intensity at each probe is measured to estimate the expression of each corresponding gene. Gene microarrays have been a useful tool in figuring out gene expression differences that differentiate one cell type from each other. In other words, we have found 'tissue signatures' of gene expression which allow us a glimpse into their working and regulation unique to their function. In my work (described in Chapter 3), I have used microarray data to study changes in gene expression with age in a part of the eye called the trabecular meshwork; identifying these signatures of aging can help us identify functions that change with age, and predispose people to glaucoma.

As with in genetic variation, microarrays for expression profiling have been rapidly supplanted by RNA-sequencing, in which mRNA molecules from a cell are first bound using polyA tags, then fragmented and sequenced. RNA-sequencing offers the ability to detect novel isoforms, as well as allele-specific expression. We now have several public expression datasets for a host of human tissues [88], including those that were previously difficult to obtain.

Studying gene expression in isolation makes it difficult to ascribe causation to a gene. For instance, up-regulation of a gene in a phenotype could either imply that changes in expression of the gene led to the phenotype, or that changes in the phenotype have led to modified regulation of the gene. A new paradigm is to merge genotype data with expression information. Genetic variants (to which causality can be attributed, since they are unaffected by environment) contributing to variation in gene expression are christened expression quantitative trait loci, or eQTLs. Currently, most studies are well-powered to detect cis-eQTLs of stronger effect sizes, while a majority of trans-eQTLs remain unknown. Combining eQTL information with QTLs can lead us to paths of causation in genetics—our ultimate goal. In Chapter 4, I integrate genetic data with expression data from RNA-Sequencing to narrow in on causal variants for aerobic capacity.

1.4 Epigenetic Information

In 2012, the ENCODE project released a rich resource of 'epigenetic' information from diverse human tissues [36]—which means information beyond anything encoded in the genome sequence. This includes regions of open chromatin, histone modifications, and regions of methylated DNA, all of which could be very loosely described as 'functional', by virtue of having some biochemical activity. For the first time, we could study the regulatory architecture of tissues on a genome-wide scale. Though the initial release was met with both fanfare and criticism [51], the ENCODE project and similar follow-ups have been invaluable resources for the scientific community for interpreting previous GWAS and sequencing results. GWAS hits were seen to be enriched in regulatory regions of the genome, and GWAS signals for certain phenotypes were seen to be enriched in regulatory features specific to the tissue involved in that phenotype. These datasets, along with gene expression information, have also helped us identify the tissue of origin of many diseases in which it wasn't clear [39].

Combining expression information with epigenomic information is heralding a new wave of integrative analysis to understand genome function in both normal and pathological samples. While the triplet code of protein-coding regions has been deciphered, the ongoing challenge is to decipher the regulatory code underlying gene regulation.

1.5 Our current paradigms about complex traits, and how we move forward

The availability of the complete genomes of organisms has shifted research towards global perspectives on life processes—to study the role of all genes or all proteins at once. The 21st century biology is likely to focus on the study of entire biological systems, by attempting to understand how individual parts collaborate to create the whole [79].

The deluge of trait-associated loci has shifted the focus from discovery to interpretation and functional validation, in what is being dubbed the 'post-GWAS era' [44]. As the authors of the above paper say, "the availability of data is not synonymous with the presence of meaning". The number of GWAS studies has not been matched by a similar number of follow-up functional studies, which means that we are often left with large lists of 'trait-associated loci' with a limited understanding of their biological contribution to the trait; I argue that the bottleneck has shifted back to the wet bench, and calls for more creative analyses to identify functional variants. We have seen some elegant studies moving from genetic variation to gene regulation to changes in expression into changes in function to changes in phenotype [136, 28]. We are now able to obtain expression levels and epigenetic information from single cells and better understand the heterogeneity within a single tissue, and mosaicism. While long-read sequencing approaches are currently not cost-effective for most groups, these are likely to be the approach of choice in the near future, and will enable us to chip away further on structural variations, which have historically been hard to study because reads from these regions have been harder to align to their reference genomes.

Has the Human Genome Project moved us towards our goal of personalized medicine? We now use the combined effect sizes from loci across the genome to calculate a polygenic risk score (PRS) for each person. However, these PRSs often increase the odds of disease by small amounts. New models to explain the small effect sizes in GWAS suggest that causality may be distributed between a large number of genes [17]; it remains to be seen if a further increase in sample sizes in GWAS and sequencing studies will lead to a convergence in genetic signal (indicating core functions implicated in a disease) or a divergence (indicating widespread causality). Furthermore, PRSs are being shown to ancestry-specific [92, 31], which would require reference samples for matched groups before we apply them to a sample.

Population cohorts are becoming larger and deeper, and we are gaining access not just to larger datasets, but also to deeper phenotype data. In 2017, the UK BioBank released the genotypes and phenotypes of 488,377 individuals, with linked electronic health record (EHR) information available for these samples [20]. Similar cohorts from other populations are also becoming available, leading to the next wave of genome-wide population-based analysis. EHR-linked samples may also allow us access to longitudinal phenotype information for samples, which may help us define better phenotypes, instead of the single-time-point information we currently use. Another source of such information will be wearable devices, which, when combined with genetic data, can be a rich source of data to mine not merely for disease status, but variations in phenotypes within the same individual.

Another much-needed (and some might say much-delayed) advance in complex trait genetics will be the increase in diversity of genomic cohorts. We will soon genetic data from large populations across the world [[21, 69, 54]. Not only is this essential for personalized medicine and disease risk scores in these populations, it adds to our understanding of how genetic background shapes the effects of risk variants.

There have been claims that the abundance of genomic data in populations will make model organism research obsolete in the study of human complex trait genetics [137]. However, model organisms offer us the only opportunity to study organism-level phenotypic consequences of mutations—mutations predicted to be deleterious in single cells are often not pathogenic in the organism (though the use of organoids may partially solve this).

In the following chapters, I will describe the approaches I have used to identify genomic signals. The tools I describe are likely to be a key piece in solving the puzzle of complex traits in the 21st century.

CHAPTER II

Exome Sequencing to Identify Rare Causal Variants in Pedigrees with Bipolar Disorder

2.1 Background

Bipolar disorder (BPD) is a severe mood disorder characterized by alternating episodes of mania and depression. The lifetime rate of BPD is around 1% in most countries [142]. The etiology of BPD is largely unknown, and there is no reliable biomarker. As a result our ability to understand and treat BPD remains at its early stages. Family, twin, and adoption studies have shown that BPD has a strong genetic component [95, 127]. However, candidate-gene studies have not yielded consistently replicable findings [26, 63]. Linkage analyses have highlighted multiple genomic regions, although the results vary [10, 97, 124];, and significant challenges remain before this approach can lead to the discovery of causal genes. Meanwhile, genome-wide association studies (GWAS) and meta-analyses have identified common variants significantly associated with BPD, implicating approximately 10 genes [101, 41, 111]. More recently, several groups have launched population- and family-based sequencing studies to investigate the potential involvement of rare variants in psychiatric disorders [102, 113, 144], including BPD [5, 30, 45, 50]. Many such sequencing-based studies are currently ongoing [129].

Given the complex genetic and phenotypic heterogeneity of BPD, family-based approaches are particularly suited to detect rare, high-penetrance causal variants that act in individual families [110]. In families with multiple cases across multiple generations, for instance, the contribution of environmental factors may be less complex than in sporadic cases, and some of such families may transmit one or a few high-impact coding variants that alter the function of a small number of pathways. To examine this scenario requires unbiased discovery of rare variants by sequencing family samples. In this study we analyzed 34 multiplex and multi-generational families by exome sequencing, aiming to identify potentially high-penetrance exomic exonic variants that increase BPD risk in each family. While the list of candidate genes may vary across families, we sought to identify convergence in functional perturbations across families that could shed light on shared biological mechanisms—even if the genes involved are highly heterogeneous among families. Thus, this approach offers a potential advantage over traditional linkage analysis, which requires accumulation of positional signals across families and has reduced power in the presence of locus heterogeneity. Whereas in this study, the strategy is to use rare variant data to first identify plausible candidate genes within each family, and then accrue pathway-level information across families.

2.2 Methods

2.2.1 Pedigrees and DNA samples

We identified 34 families for this study from the NIMH Bipolar Disorder Pedigree Catalog maintained by the NIMH Center for Collaborative Studies on Mental Disorders. The inclusion criteria were the families (1) with multiple affected individuals in multiple generations and (2) containing at least one first cousin pair or more distantly related pairs who are both affected with Bipolar Disorder Type 1 (BP1). We obtained lymphoblastoid cell line-derived genomic DNA for 344 subjects representing all members in these families for whom DNA samples were available (195 affected with BPD or related mood disorders, and 149 unaffected). Diagnoses for all subjects were obtained from the "Dx" field of the NIMH data file "bp_ped_6_02.csv" downloaded on June 17, 2013 from NIMH Data Repository Bipolar Disorder Dist. 7.0. On average there were 10 subjects from each pedigree, with a range of 3-22. The number of subjects per family is shown in Table 2.1. These pedigrees contained both BPD cases and those diagnosed with other psychiatric phenotypes: in total, there were 125 subjects with BP1, 18 with Bipolar Disorder Type 2 (BP2), 35 with Recurrent Unipolar Depressive Disorder (RUDD), 11 with Schizoaffective Bipolar Disorder, and 6 subjects with unclassified mental disorders. The remaining 149 subjects were not diagnosed with any psychiatric disorder.

2.2.2 ExomeChip genotyping

To confirm familial relatedness and sample quality we performed genotyping on all 344 DNA samples using the Illumina HumanExome BeadChips (version 1) at the University of Michigan Sequencing Core. Quality control of the genotype calls was performed using PLINK version 1.07 [112]. Genotypes were obtained for 247,870 SNPs. All samples had a genotype missingness rate of <1%. Across 90 samples, 247,046 (99.7%) SNPs had per-SNP missing rate of < 5%; thus those with missing rate >5% were removed from further analysis. Three samples in three different families failed sex check (estimated by *PLINK* using heterozygosity of the X chromosome) and were removed from further analysis, leaving a dataset of 341 samples.

2.2.3 Verifying relatedness using genotype data

We adopted two strategies to assess the level of relatedness between pairs of individuals. First, we used BEAGLE fastIBD [19] to identify genomic segments shared identically by descent (IBD) among the 341 samples, and used the total proportion of the genome in IBD as a measure of relatedness. Second, we calculated the proportions of the genome in IBD of 1 or 2 (i.e., Z score of 2 or 1) between each pair with the –genome function in PLINK, using 4,860 “Grid SNPs”, incorporated in the ExomeChip to cover the genome with regularly spaced common variants. We found that the level of IBD as calculated in PLINK is similar to that from BEAGLE, with a Pearson's correlation coefficient of 0.91 over all within-family pairs (838 pairs in 34 families, counting married-in individuals). We subsequently compared the BEAGLE fastIBD scores with the expected relatedness based on the known pedigrees, and found a high level of concordance 2.1 despite a moderate bias where the observed values are often lower than the expected values. This is likely due to under-calling of smaller IBD segments using ExomeChip data.

We identified several sample pairs with unexpectedly high levels of IBD sharing: a sibling pair in Family 389313 had a value of 0.9; and an aunt-nephew pair in Family 15-00160 had a value of 0.9. We found that one sample in each pair was likely mislabeled based on his/her abnormal IBD with other members of the pedigree. We removed the two samples in question, leaving a dataset of 339. Besides these, several first-cousin pairs in Family 38-9064 had IBD scores in the range of 0.38-0.53, much higher than the expected Z1 + Z2 of 0.25 for first cousins. The related parents of the cousins, who are siblings, showed the expected level of IBD for sib pairs, thus the result could be due to cryptic relatedness of one of the married-in members. However, this cannot be tested as the DNA is not available for the married-in members in this family. A similar scenario of cryptic relatedness is suspected for a married-in member in Family 10101, who had IBD of 0.11-0.39 with other members of the family. These samples were not removed because they were not sequenced and had little impact on the linkage analyses.

2.2.4 Exome sequencing and variant calling

For exome sequencing we selected 2-4 BP1 cases from each family representing at least one pair of first cousins or more distantly related pairs, and sometimes included an additional affected sibling. The reason to focus on cousins or more distant relatives is that they have lower levels of allele sharing due to IBD than siblings. In all, 90 samples were sequenced. Exome capture was performed using the Roche Nimblegen SeqCap EZ Human Exome Library v3.0, targeting a total of 64 Mb of the genome. Paired-end, 100-base sequencing data were collected using the Illumina HiSeq2000 System. Exome capture and sequencing were performed in two batches. All 90 samples were sequenced in the first batch. Of these, 34 samples with the lowest coverage were re-captured using the same capture kit, and sequenced in the second batch to increase coverage. The final read depth across 90 samples ranged from 20X to 80X, with a median of 51X. Among the called variant sites, 94% have at least 10x median coverage across all samples.

We aligned sequence reads to the human reference genome GrCh37 using BWA version 0.5.9 [84], removed duplicate read pairs using PICARD version 1.74 [18], and merged reads from two batches for samples that were sequenced in both. We performed multi-sample joint calling of variants using GATK version 3.1 [96] in a 300-exome pool, including the 90 BPD samples and 210 other, non-psychiatric samples undergoing exome sequencing in concurrent studies in our group. We applied the GATK VQSR filter to remove low-quality variants, resulting in 470,021 on-target pass-QC variants (441,278 SNPs and 28,743 indels). The transition to transversion ratio is 2.4. For these pass-QC variant sites we defined low-quality genotypes by a genotype quality ("GQ") of < 10, and marked them as missing. Variant annotation was retrieved by using ANNOVAR [140] on July 27, 2014.

As a technical validation of the single nucleotide variants (SNVs) discovered by sequencing we compared the called genotypes in the 90 samples with those from the ExomeChip data, and found a high level of concordance. As shown in Table 2.2, the average concordance rate is 98.6% among the 20,484 sites both called by sequencing and genotyped by the ExomeChip, even when the missing genotypes are counted as discordant. Ignoring genotypes missing in either platform led to an even higher concordance rate: 99.6%. Since the homozygous reference genotype is by far the most abundant genotype in this panel of rare variants, we also estimated the concordance for heterozygous or alternative homozygous genotypes called in either platform. Among these, the concordance is 98.2% if counting missing genotypes as errors, 99.4% when ignoring missing calls. These results show that our downstream analysis was based on very high-quality genotype calls.

2.2.5 Variant filtering

The goal of variant filtering is to identify functionally damaging variants that are rare in the general population and shared among BPD cases in each family. First, to focus on rare variants we removed those with a minor allele frequency (MAF) >1% in the European subset of the 1000 Genomes Project (514 samples) or in the Exome Sequencing Project (4,300 individuals of European ancestry) (Nhlbi Exome Sequencing Project (Esp), 2011) . Second, we selected variants that are nonsense, residing within 2bp of a splice junction, causing frameshift, or missense variants with predicted damaging effect according to either SIFT [103] or PolyPhen-2 [2]. For PolyPhen-2, only variants classified as "probably damaging" were kept. Third, since some false positive variant calls might arise due to the specific sequencing and variant calling processes adopted in each laboratory, we identified variants that are rare in public databases but are common (MAF>5%) in our "local" exome database, consisting of 210 non-BPD exomes that we sequenced and analyzed in other, concurrent studies using the same procedures. The numbers of variants (SNVs and indels) remaining at each of the three stages described above are shown in Figure 2.1.

Next, we proceeded to filter variants within each family by ignoring variant sites that are monomorphic in that family (as they were identified in other families in the joint calling). We then kept variants shared among all the sequenced cases of the family. Note that in some families there are other BP1 cases not yet sequenced and if they were, could further reduce the number of shared variants. Since there were missing genotype calls due to low coverage, we allowed a variant to be considered as shared by the sequenced cases in a given family if at least half the sequenced cases had a called heterozygous genotype containing the rare allele, and the remaining sequenced cases, despite having apparently "missing" genotype, met certain criteria supporting the presence of the rare allele. The criteria are: having at least one sequencing read supporting the alternative allele and the alternative allele frequency being 1/6 or higher. This increased the number of shared variants by about a third.

After obtaining the list of shared variants for each family, we further narrowed each list by segregation patterns among all the genotyped members, because exome sequencing was done only on 90 individuals; yet all members with DNA were genotyped. A perfectly segregating variant would be shared among all cases but absent in all unaffected samples. In order to allow for imperfect segregation due to reduced penetrance we used each family's linkage analysis LOD scores calculated from the genotype data (described below) as a surrogate measure of segregation patterns along the genome. For each family, we included variants in regions that have LOD scores greater than 0.

2.2.6 Segregation patterns by linkage analysis

The goal of linkage analysis in this step is to find genomic regions with higher, albeit often imperfect, segregation patterns within each family, and use these regions as a segregation filter for that family.

Phenotypic categories ("models"): To address phenotypic heterogeneity we ran linkage analysis separately for three diagnostic models of BPD commonly used in linkage studies. First, the "BP1 model" considers only BP1 individuals as affected. Second, the "BP2 model" considers either BP1 or BP2 individuals as affected. Third, the "Bipolar Spectrum", or "BPS model", considers BP1, BP2, or RUDD individuals as affected. In all three models, only individuals without any mood disorder were classified as unaffected, and the individuals with some mood disorders but not in the BP1, BP2, or BPS categories were classified into the "unknown" category. This way, individuals with the Schizoaffective Bipolar Disorder and unclassified mental disorders were always coded as "unknown". Parametric linkage analysis: We performed parametric linkage analysis using *MERLIN* version 1.1.2 [1], applied on the set of 4,566 autosomal "Grid SNPs" on the ExomeChip. We adopted the dominant model by setting penetrance values of 0.0001, 0.7 and 0.7 for the three genotypes, with the homozygous rare allele genotype and the heterozygote genotypes having penetrance of 0.7.

2.2.7 Co-expression in specific brain regions and developmental periods

To study co-expression patterns within any gene lists we used microarray-based gene expression data from the BrainSpan Project [98]. This project measured the human brain transcriptome at 27 stages of development, from 12 post-conception weeks to 40 years, and for 26 regions of the brain. We downloaded the file "exon_array_-matrix.csv.zip", containing normalized data for 493 samples over 17,604 genes (Miller et al., 2014), from BrainSpan.org on November 1, 2014. Following the procedures described before [53] we selected 15 brain regions and divided them into four broad anatomical categories 2.1: sub-cortical regions, sensory-motor regions, frontal cortex, and temporal-parietal cortex. We also divided the time points into three developmental periods: fetal, childhood (from early infancy to late childhood), and adulthood (from adolescence to late adulthood). This led to 12 experimental groups, containing

15-62 samples each. For each of the 12 groups we calculated the Pearson correlation coefficient (r) for all gene pairs across samples within the group. This within-group analysis ensures that the gene co-expression networks are not unduly inflated by the large-scale concerted changes across broad regions and distinct developmental stages. Gene pairs with r 0.8 were considered “interacting” in the co-expression network.

To determine if the number of interacting gene pairs among the putative BPD genes is more than other “random” gene sets, we compared the number of edges in our gene list with the number of edges in a ’control’ set comprising genes with non-damaging variants, but passing every other filter. Thus the control genes are those in the same co-segregation/linkage regions but containing non-damaging variants, chosen to contrast with the damaging variants in the same regions. We used the non-parametric Kolmogorov-Smirnoff (KS) test to compare the two distributions. We chose a conservative threshold of P-value<0.004, to reach overall P<0.05 after correcting for testing 12 groups. This strategy avoided the pitfall of building a null distribution of the number of edges by repeated sub-sampling of the control gene sets, as these sub-samples share many of the same genes and are not independent, leading to a null distribution with an improperly small variance and vastly inflated significance.

2.2.8 Pathway analysis

We applied the Ingenuity Pathway Analysis (IPA, www.qiagen.com/ingenuity, QIAGEN, Redwood City, CA, run on June 12, 2015) and DAVID [66] to assess pathway enrichment in the BPD genes. These tools identify biological functions or pathways that are enriched in any user-defined gene list. In IPA we used the default settings and tested 653 pathways defined as “canonical metabolic and signaling pathways”. DAVID compares an input gene set with a “baseline” gene set for enrichment of functions. To control for potential systematic bias introduced at various stages of variant filtering, we cannot designate all known genes, or a random subset, as the “baseline”. Rather we need to match the baseline genes to the final set of BPD genes as closely as possible. To do so, we designated the baseline genes as those represented by variants passing all but the last filter, thus having met the same series of criteria for population frequency, functional prediction, local exome control, and sharing among the sequenced cases, before the final, LOD score-based segregation filter.

2.2.9 External databases and gene lists

Autism genes: We downloaded a list of 859 autism-related genes from the manually curated SFARI database (2017) on March 31, 2017. This database includes high-confidence autism genes categorized into rare, syndromic, association and functional categories, based on the evidence connecting these genes to autism.

Schizophrenia genes: We obtained a list of 338 gene associated with schizophrenia from Supplementary Table S3 of [119]. These genes were those in the vicinity of 108 SCZ-related loci (as defined by Ripke et al) obtained from the largest GWAS of SCZ, with 36,989 cases and 113,075 controls.

Major depression-related genes: We obtained a list of 69 gene associated with depression from Table 1 of Wray et al, 2017, who performed a meta-analysis of 130,664 cases and 330,470 controls.

BRIDGES data: The Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) study sequenced the genomes of unrelated individuals (1,789 cases with BPD and 1,927 controls) of European ancestry to a mean coverage of 9.6X. We obtained the SKAT-O burden test p-values and global ranks for 37,875 genes from an interim freeze in February 2017.

2.3 Results

2.3.1 Variant filtering and discovery of potential BPD genes

We performed exome sequencing for 90 BP1 cases in 34 families, choosing in each family at least one pair of first cousins or more distant relatives who are both affected. After variant calling we processed the on-target bi-allelic variants by a series of filters 2.2. By excluding common variants in the general population we removed $\tilde{3}$5% of the observed variant sites. We then kept only those likely to have a functional impact: introducing a stop codon, creating a frameshift, residing near a splice junction, or being missense and predicted as damaging by either of two function annotation tools: SIFT and PolyPhen-2. This step removed 87% of variants remaining from the previous step. Third, in order to control for laboratory-specific false positives in the sequencing and variant calling procedures we excluded variants found in high frequencies (MAF>5%) in 210 non-psychiatric subjects that were sequenced by the same protocols in our laboratory and included in the joint variant calling. Next, we analyzed each family to identify variants shared among sequenced cases in that family. Lastly, we further selected variants in regions of increased sharing among cases, using family-wise LOD>0 as the cutoff. The three phenotypic models (see Methods) led to slightly different LOD>0 regions. Among the 34 families, 31 have at least one region passing this threshold with at least one of the models.

2.3.2 Regions of increased sharing: Family-wise signals of suggestive linkage

When linkage scores were summed across families, no region reached genome-wide significance (LOD = 3), reflecting potential locus heterogeneity across families and the limited power to overcome such heterogeneity given our sample size. Per-family linkage analysis did not reveal regions reaching LOD = 3 in any family. However, some families showed regions of suggestive linkage (LOD score >= 2.2) on chromosomes 4, 12, and 13. Specifically, Family 11107 had LOD = 2.63 in 12q21-23 for both the BPS and the BP2 models, and LOD = 2.39 in 12p12.1 for the BP2 model. Family 15-00118 had LOD = 2.6 in 13q32 for the BPS model. Family 11150 had a LOD score of 2.69 in 4q35.2. These regions did not overlap with those identified in a recent linkage study of 972 BPD pedigrees. However, some of the regions have been implicated in other studies, as described below.

Region 13q32 was associated with BPD and schizophrenia independently in several previous studies [33]. Interestingly, the studies implicating 13q32 used the Bipolar Spectrum classification, consistent with this region being observed in the BPS model in our study. The 12q23-24 region reached genome-wide significance in two previous genome scans [38, 128] and was also implicated in a linkage analysis of unipolar disorder [29]. Since this region is significant in our BPS model, it may underlie both the unipolar and the bipolar disorder. The 12p12 region has shown suggestive linkage with BPD in Ashkenazi Jewish families [9]. The region 4q35 has been previously implicated in other linkage studies of BPD and schizoaffective disorder [**?** 108].

Although these were the top linkage regions, they did not contribute to the final list of potential BPD genes in those families, because earlier filters for damaging variants shared among the sequenced cases had left no variant in the family-wise linkage region. It is possible that we are missing variants within these linkage regions segregating in other affected members that havent been sequenced, or it is possible that we missed regulatory variants that were not recovered by exome sequencing.

These filtering steps led to rare-damaging-shared variants in 310, 282 and 239 genes with the BP1, BP2 and BPS models, respectively. The three lists of genes have significant overlap, with a union of 336 genes. The list of genes implicated in each family is shown in Table S3. We find an average of 8-9 genes per family (range: 0-18).

If some of the pedigrees shared a genetic cause of BPD, this could be seen in a convergence in causal genes across families. However, most of our genes were found in only one pedigree; only 6 genes have damaging mutations in two pedigrees in at least one model (HLA-C, SYNJ2, GNRHR, EVPL and RFX1 in BP1 model; EVPL, HLA-C, RFX1 and SYNJ2 in BP2; EVPL and MORN1 in the BPS model).

2.3.3 Overlap with prior genetic findings

Between the 336 genes identified in this study and the 353 schizophrenia genes identified in a GWAS meta-analysis [119], seven genes (ACTR5, SLC32A1, ALDOA, SMG6, TBC1D5, ATXN7 and PCDHA4) appeared in both; however this rate of overlap is not significantly higher than expectation. No gene overlapped with the 69 genes associated with the major depressive disorder. Of the 859 autism genes from the SFARI database , 25 appeared in our list 2.3, 2.1 fold of the expected number of overlap ($P < 0.006$). This includes CREBBP, (CREB Binding Protein), which is a circadian rhythm gene. Recently, the BRIDGES Consortium performed whole-genome sequencing and comparisons of rare variant burden between 1,789 BP1 cases and 1,927 controls. In unpublished SKAT-O test results, 14 of our 336 genes had P < 0.05, which is within the expectation by chance. Rank-based analysis found no departure from a random distribution of our 336 genes (Kolmogorov-Smirnoff test, not significant).

2.3.4 Expression level in the brain

We obtained normalized gene expression read count data from the Genotype-Tissue Expression Project (GTEx) to test if our list of 336 genes tended to be highly expressed in the brain. We obtained the distribution of read counts for our gene list in 13 regions of the brain from the GTEx portal on November 7, 2016, and compared this distribution with that for our control gene list (1,955 genes with non-damaging variants in our dataset) using the non-parametric Kolmogorov-Smirnoff test. We didn't find any significant difference between the two distributions (results not shown).

2.3.5 Enrichment analysis of gene networks

To assess if the BPD genes identified in this study are more likely to interact with each other than a properly matched set of baseline genes, we analyzed (1) protein-protein interaction data and (2) gene co-expression data.

Protein interaction network: We used the GeneMANIA database of interacting proteins [141] to assess if there were more interacting gene pairs in our list than expected by chance. For each disease model, we compared the number of interactions for each gene in the list of genes in the candidate gene list with the number of interactions in a control gene list obtained by filtering all non-damaging variants passing all our other filters. We do not see an increased number of interacting gene pairs in our candidate list compared to our control list of genes using the BP1 (Figure 2.4), or those from using the BP2, and BPS models (not shown).

Co-expression network: Previous studies of autism [143] have used co-expression networks to show that genes associated with autism converge upon biological pathways in layer 5/6 cortical projection neurons in fetal developmental time points. Because of the enrichment of autism-related genes in our candidates, we applied a similar method to study if BPD-associated genes converge upon pathways in specific regions of the brain, and at the same time points. We downloaded expression values from the publicly available microarray data from BrainSpan (see Methods), and used the non-parametric Kolmogorov-Smirnoff (KS) test to compare the distribution of co-expression between our list of candidate genes and our list of control genes. Figure 2.3 shows that after correcting for multiple-testing across 12 groups, we see no significant enrichment of co-expression across all 12 groups. However, the time period of infancy to childhood and in the sub-cortical region, shows the greatest co-expression of our candidate genes, which is nominally significant. The results for the BP2 and BPS models were similar (not shown).

2.3.6 Pathway analysis

In IPA analyses, of the 653 tested pathways tested, circadian rhythm signaling was the top pathway in all three models, though it did not cross the significant threshold using FDR 2.3. The functional link between circadian rhythms and BPD has been reported in many previous studies [15, 93], including our own [85]. This result is driven by four genes, ATF4 (activating transcription factor 4), CREB3L4 (cAMP responsive element binding protein 3-like 4), CREBBP (CREB binding protein), and PER3 (period circadian clock 3), that are implicated in six families. DAVID did not identify any pathway with significant enrichment among Gene Ontology, KEGG and InterPro pathways.

2.4 Discussion

In this study, we used exome sequencing to discover coding variants and applied multi-stage filtering to find rare damaging variants that segregate with BPD in each of 34 multiplex families. After finding the genes carrying such potential risk variants in individual pedigrees we sought to find convergent signals across pedigrees. This strategy is motivated by the recognition that linkage analysis has reduced power in cases of locus heterogeneity, due to its ineffective accrual of positional signals across families. As an alternative, our approach tests the genetic model that there are shared functional perturbations across families, and that such evidence can accumulate at pathway level even if different genes are implicated in different families.

As expected, our linkage analysis did not find regions of significance across families, consistent with the small sample size used here and the hypothesis of locus heterogeneity. However, some regions with suggestive linkage in individual families have been previously implicated in BPD. Region 13q32 was associated with BPD and schizophrenia independently in several previous studies, including different pedigree-based analyses [33]. Interestingly, the studies implicating 13q32 used the Bipolar Spectrum classification, consistent with this region being observed in the BPS model in our study. The 12q23-24 region reached genome-wide significance in two genome scans [38, 128], and was also implicated in a linkage analysis of unipolar disorder [29]. Since this region is significant in our BPS model, it is plausible that it underlies both unipolar and bipolar disorder.

Our variant filtering protocol led to lists of 329, 286 and 266 genes using three nested definitions of cases. Most families had multiple genes implicated (average 7.8-9.5 genes/family across the three models). Our list of genes includes some that recur in two of the 34 families (HLA-C and EVPL); and a subset of these has been previously linked to BPD or SCZ in genetic studies. While these observations are consistent with the scenario that some of the pedigrees have an oligogenic form of inheritance, with variants in multiple genes jointly leading to the high BPD risk in each family, our results lack the power to further prioritize the candidates.

At the pathway level, our BPD gene candidates are enriched for those involved in circadian rhythm signaling pathway and together they affected six of the 34 pedigrees. Several genetic studies have revealed mutations in clock genes in BPD patients [94, 104, 125]. Sleep disruption and circadian dysregulation are well-known comorbid factors of BPD. Nearly all patients with BPD have severely disrupted circadian rhythms, which in turn influence sleep cycles [93]: a reduced need for sleep is a common symptom of the manic episodes, while increased sleep is often seen in the depressive episodes [59]. Moreover, regulation of sleep patterns has often brought mood stabilizing effects in BPD patients [77].

A number of family-based sequencing studies in BPD have been published, but with limited convergence in results [30, 45, 50, 115]. The lack of convergent signals in our pedigrees further highlights this pattern, suggesting that rare variants also follow complex patterns of inheritance in BPD. An overlap with autism-related genes points to an overlap in genetic architecture between different neurological disorders.

Our study design relied on unbiased variant discovery, rigorous filtering, and a systematic evaluation of functional signals accumulated over the 34 families, covering a wide range of bioinformatics data sources and carefully constructing the null distributions and control gene sets. After successfully executing this plan we uncovered enrichment signals for circadian genes and autism genes, although the statistical effect size were moderate, and most functional themes were not over represented. For the narrow goal of finding pseudo-Mendelian families where very high-penetrance variants are segregating, our results were inconclusive as we lack the crucial support of recurrence of the same genes in multiple families. Large collections of such "heavily loaded" families will be needed to meet this goal, meanwhile it remains possible that very few of the familial BPD cases involve the strong-acting alleles in a small number of genes. Combined analysis of both population-based and family-based data, preferably with pedigrees embedded in population samples, has received increasing appreciation and will likely drive the next phase of research for this severe psychiatric disease.

Figures

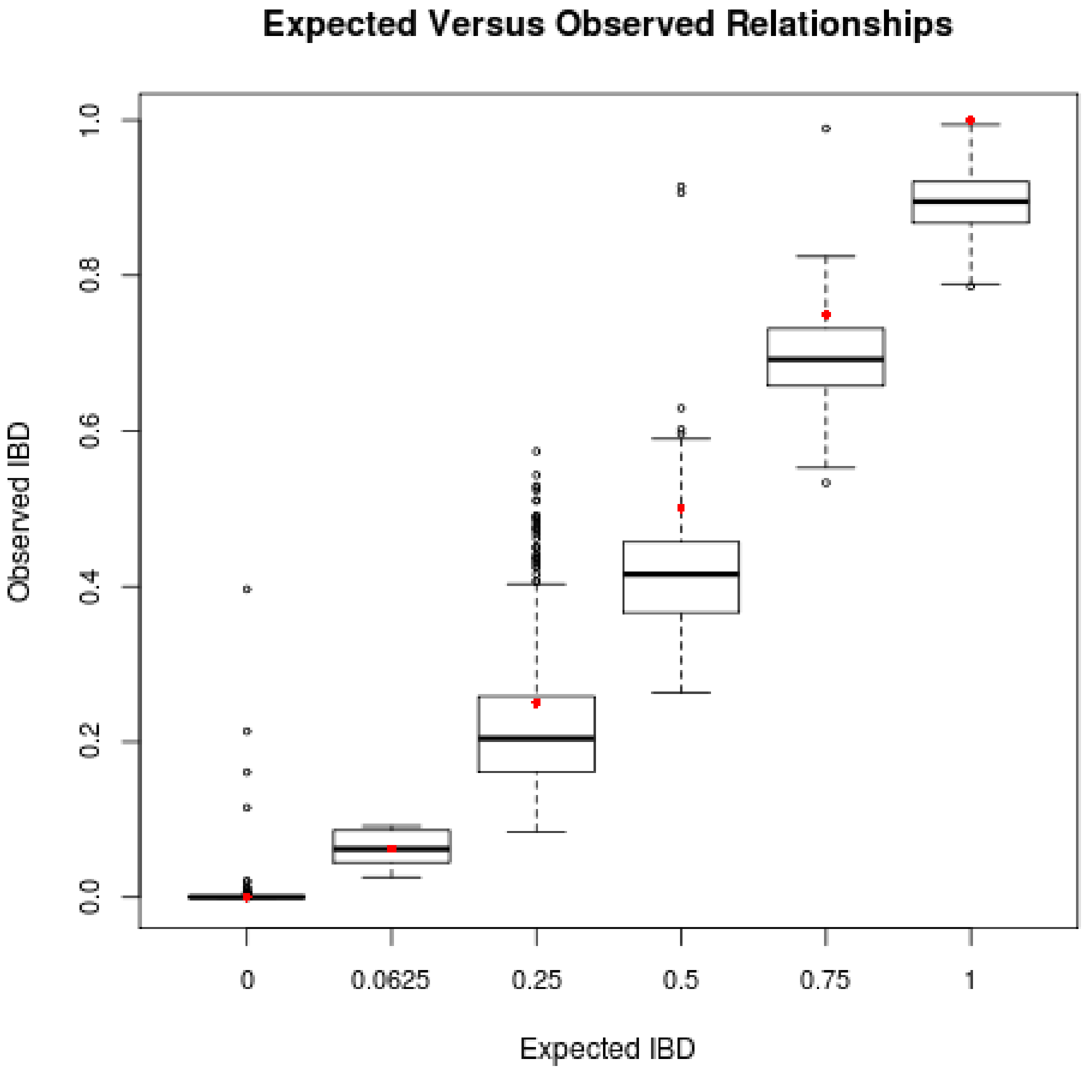

Figure 2.1: **Expected versus observed relationships** Comparison of the expected genetic relatedness based on known pedigrees and the inferred relatedness based on sequencing data. The x and y axes show the expected and the inferred IBD, respectively, for different degrees of relatedness. The red dots indicated the expected value.

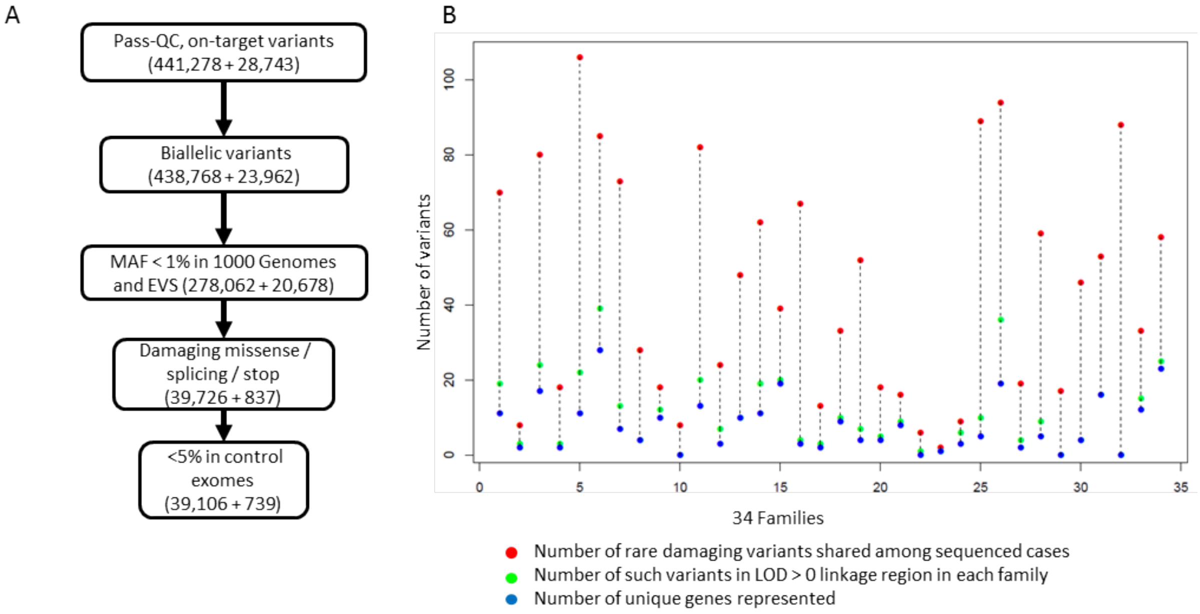

Figure 2.2: **Variant discovery and analysis.** The number of variants left after various filtering steps in each of 34 pedigrees (x axis).

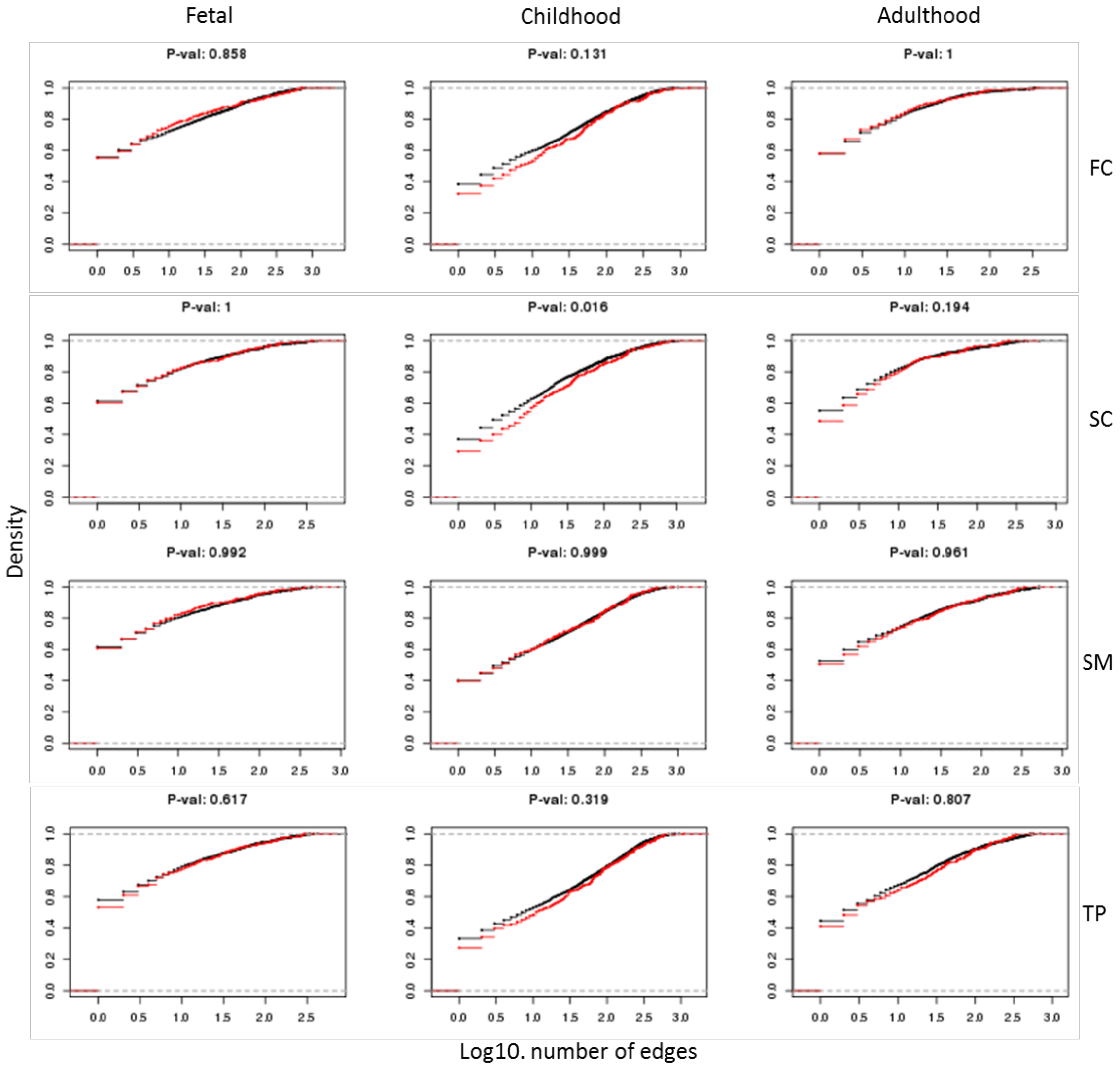

Figure 2.3: **Kolmogorov-Smirnoff test for gene co-expression between our candidate gene list and control genes** Comparing the co-expression distribution within our set of damaging genes compared to the co-expression distribution for a control set of non-damaging genes for 4 brain regions across 3 time points. The plot below shows the number of edges for each gene in our candidate gene list (red) and in our control gene list (black.) A graph shifted towards the right indicates a greater number of pairwise interactions, as defined by a Pearson correlation coefficient of ¿ 0.8.

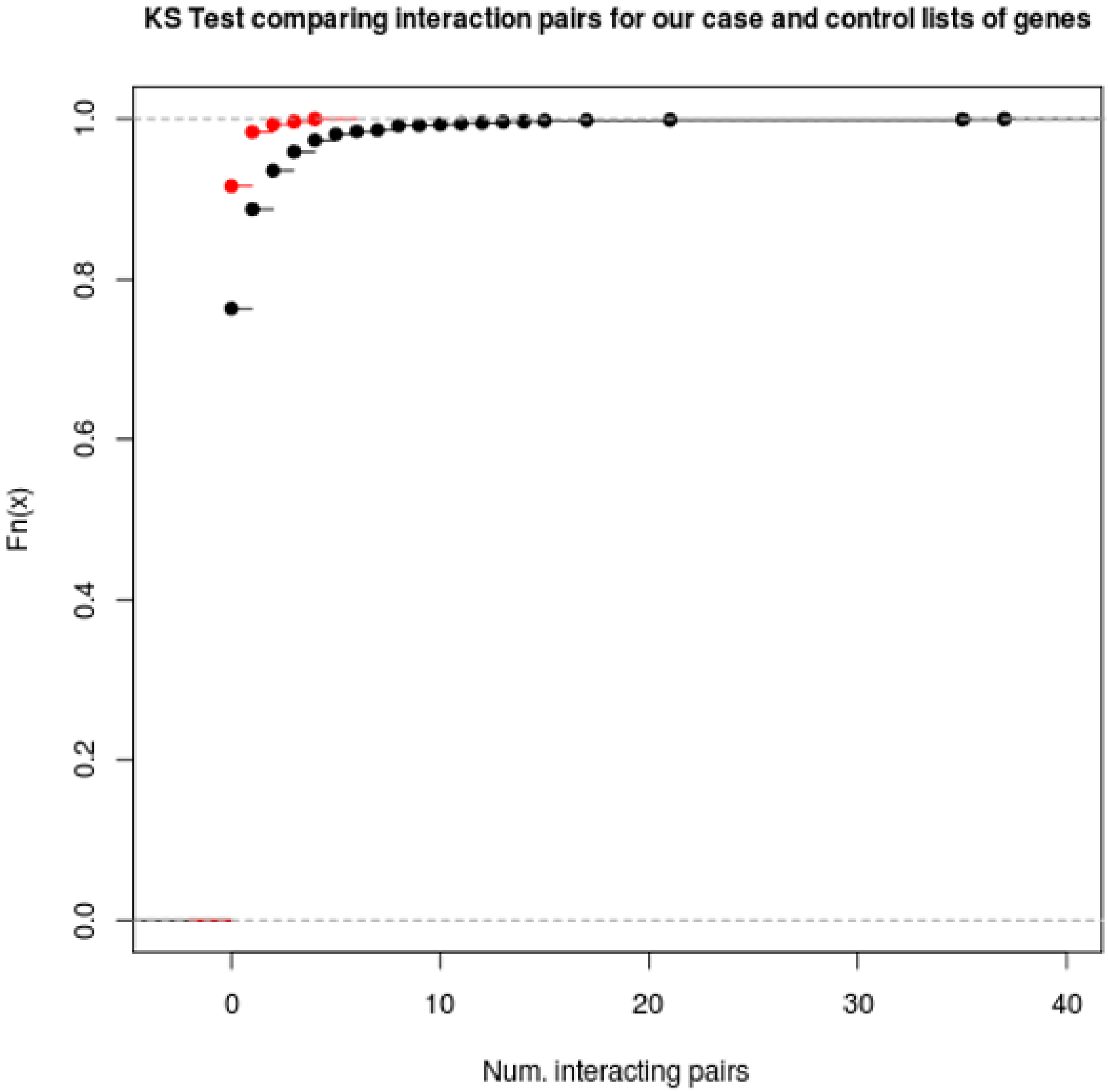

Figure 2.4: **Kolmogorov-Smirnoff test for protein-protein interaction between our candidate gene list and control genes** This plot displays the cdf of the number of interactions for each gene in a list of genes. The number of interacting gene pairs (as defined by GeneMania) that were found in our list of 310 genes (shown in red) was not significantly higher than the number of interacting gene pairs found for each gene in our list of control genes (A list with a significantly higher number of interactions would be shifted to the right). A Kolmogorov-Smirnov test testing for a greater number of interactions in our case list returned a P-value of 1.

Tables

Family ID	#Geno--typed	#Sequ--enced	#Cases (BP1)	#Cases (BP2)	#Cases (BPS)	Comments
10101	8	2	2	3	6	
10102	14	4	6	7	11	
11101	12	2	2	2	5	
11107	13	4	5	7	8	
11108	7	2	4	5	7	
11122	7	2	3	3	4	
11130	7	2	3	3	5	
11150	15	3	5	5	7	
11153	8	3	3	3	6	
11156	19	3	6	6	8	
11164	7	2	3	3	4	
12330	9	3	4	6	7	
13101	13	2	2	2	5	
13126	10	2	3	4	8	
13139	7	2	4	4	5	
15-00113	11	2	3	4	5	
15-00118	13	3	5	6	9	
15-00120	10	3	3	4	6	
15-00136	10	2	2	4	5	One sample removed (gender mismatch)
15-00160	14	4	8	8	8	Two samples removed (gender/relatedness)
15-00164	9	3	3	3	6	
15-00181	14	4	4	6	9	
15-00207	10	4	4	4	4	
15-00502	8	3	3	3	3	
22-1003	9	2	3	3	6	
22-1004	5	2	3	3	3	
25-1002	5	4	5	5	5	
26-1020	11	2	7	8	8	
38-8031	17	3	4	4	5	
38-8042	12	2	2	3	3	
38-9056	5	2	3	3	4	
38-9064	14	2	3	3	5	
38-9313	3	2	3	3	3	One sample removed (relatedness check)
38-9326	8	2	2	2	2	One sample removed (gender mismatch)

Table 2.1: **Number of sequenced and genotyped samples in 34 families**

		Exome Chip				
		N_0/0	**N_0/1**	**N_1/1**	**N_./.**	**Concordance**
Exome Sequencing	**N_0/0**	1,438,297	2,216	1,573	3,122	0.995
	N_0/1	653	244,435	525	767	0.992
	N_1/1	1,276	1,758	135,839	566	0.974
	N_./.	9,087	1,982	1,420	44	
	Concordance	0.992	0.976	0.975		0.986

Table 2.2: **Genotype concordance between ExomeChip and exome sequencing results in 90 samples.**

			Number of samples for each developmental period		
Region Code	**ID**	**Tissue Name**	**Fetal**	**Childhood**	**Adulthood**
SC_a	STR	Striatum	14	5	5
	MD	Mediodorsal nucleus of thalamus	9	6	5
	AMY	Amygdaloid complex	14	7	5
	HIP	Hippocampus	16	7	4
FC_b	DFC	Dorsolateral prefrontal cortex	17	7	5
	MFC	Anterial cingulate cortex	16	7	5
	OFC	Orbital frontal cortex	14	6	5
	VFC	Ventrolateral prefrontal cortex	15	8	5
TP_c	ITC	Inferolateral temporal cortex	12	8	5
	STC	Posterior superior temporal cortex	17	8	5
	IPC	Posteroinferior (ventral) parietal cortex	14	8	5
SM_d	A1C	Primary auditory cortex	14	7	5
	M1C	Primary motor cortex	16	7	5
	S1C	Primary somatosensory cortex	16	8	5
	V1C	Primary visual cortex	14	8	5

Table 2.3: **Number of tissue samples in gene expression data from BrainSpan, for four aggregated brain regions and three time periods.** a: Sub=cortical regions, b: Frontal Cortex, c: Temporal-parietal cortex, d: Sensory-motor regions
Fetal period: 13-26 weeks post conception.
Infancy to childhood: 4 months-8 years.
3-40 years (There was no sample for the 8-13 year interval.)

Gene Symbol	**Gene Name**
ABCA7	ATP-binding cassette sub-family A member
ACHE	Acetylcholinesterase
ATXN7	Ataxin 7
CACNA1H	Calcium voltage-gated channel subunit alpha1 H
CAMSAP2	Calmodulin regulated spectrin-associated protein family member 2
CD44	CD44
CEP290	Centrosomal protein of 290 kDa
CLTCL1	Clathrin, heavy chain like 1
CNTN3	Contactin 3
COL28A1	Collagen, type XXVIII, alpha 1
CREBBP	CREB binding protein
CUL7	Cullin 7
DCTN5	Dynactin subunit 5
DLGAP1	DLG associated protein 1
DNAH3	Dynein axonemal heavy chain 3
DOCK1	Dedicator of cytokinesis 1
GAP43	Growth associated protein 43
MYO5A	Myosin VA
P2RX5	Purinergic receptor P2X 5
PCCA	Propionyl-CoA carboxylase
PCDHA4	Protocadherin alpha-4
PHF3	PHD finger protein 3
PTK7	Protein tyrosine kinase 7
SBF1	SET binding factor 1
SLCO1B3	Solute carrier organic anion transporter family member 1B3
SMG6	SMG6
TECTA	Tectorin alpha
TBC1D5	TBC1 domain family member 5
TTN	Titin
USH2A	Usherin 2A
WWOX	WW domain containing oxidoreductase

Table 2.4: **Thirty one genes overlapping with the autism-associated genes obtained from SFARI .**

Pathway	**N1**	**N2**	**P-value**	**FDR**
Circadian Rhythm Signaling	4	33	1.5e-03	0.49
ILK Signaling	9	196	3.4e-03	0.49
Neurotrophin/TRK Signaling	8	76	6.3e-03	0.49

Table 2.5: **Top canonical pathways in the Ingenuity Pathway Analysis for 336 genes found from the union of the three phenotype models.**

CHAPTER III

Identifying Expression Signatures of Aging in the Trabecular Meshwork

3.1 Introduction

The trabecular meshwork (TM) is a tissue in the outer eye responsible for the outflow of aqueous humor from the cornea [87]. The TM is a crucial determinant of the intraocular pressure (IOP) because it lends resistance to the evacuation of aqueous humor from the eye. With increasing age, the TM shows decreasing cellularity [4] and increased stiffness [99], which could lead to increased resistance to aqueous humor outflow. This increased resistance is one of the risk factors for glaucoma, an age-related eye disease [49]. Understanding the molecular mechanisms by which this resistance changes with age can give us clues about ways to understand age-related changes in IOP.

There have been several studies on transcriptional changes with aging in human tissues such as blood, brain, muscle [121, 145, 107, 149] as well on model organisms like *C. elegans* [89], fruit flies [109], mice [82] and primates [42]. The only study on transcriptional profiling in this tissue [23] revealed 1,387 age-associated genes. However, because of their small sample size, they divided their samples into two groups—fetal and old—to identify differences between them. In our study, we attempt to identify genes whose expressions in the TM change continuously with age, and through this, understand potential mechanisms involved in aging of this tissue. Previous studies have shown that the aging process tends to be associated with small expression changes in many genes, as opposed to large changes in a few genes. The number of age-related genes that are shared across tissues remains small [48], but pathway analyses have identified some common signatures of aging. Understanding age-related expression changes in the TM can help understand tissue-specific signatures of aging, as well as estimate similarity in the aging process with other tissues.

3.2 Methods

3.2.1 Sample collection

Our samples of postmortem samples were from the University of Michigan eyebank. For most samples, the anterior segment of the eye is removed and then stored in optisol preservative. During a cornea transplant, the center part of the donor's cornea is punched out from this preserved tissue, leaving a donut of tissue that contains the TM. We collect this remnant, and from it, we remove the TM. Occasionally the eyebank collects a cornea and they are unable to use for transplant for some reason (usually unrelated to the properties of the sample). In that case, they make this "full cornea" available for research purposes. We remove the TM exactly the same way as with the surgical remnant.

3.2.2 Transcription profiling

93 RNA samples were assayed on the Affymetrix Human Genome U219 array using standard protocols of the University of Michigan sequencing core. Data was normalized using RMA [70] and summarized to probesets in the Affymetrix analysis suite, and followed by quantile normalization across the 93 samples. The array has

49,385 probesets, 2,899 of which map to multiple genes (as defined by the U219 annotation file downloaded from the Affymetrix website on April 20, 2017). We removed these multi-mapping probesets, using the remaining 46,486 probesets in downstream analysis. These probesets map to 18,560 genes, with an average of 2.5 probesets mapping to one gene. We analyzed gene expression association separately for probeset. Principal components analysis revealed no outlier samples in the top 10 principal components.

3.2.3 Regression analysis

We noticed that our oldest samples (by biological age) also had the lowest values of RIN, and that most of the older samples were obtained from whole corneas not used for transplantation (Figure 3.1). Thus, we used only the remaining 72 samples from the corneoscleral button (CSB) for our regression analysis; these were samples extracted from the remnant of the corneal tissue after it was used for a transplant. We used the *limma* [131] package in R to identify genes whose expression levels changed with age. In order to remove effects of other variables that could affect expression, we used sex, RNA Integrity Number (RIN), race and Plate ID as covariates in the regression.

We also ran other regression models using the top principal components as covariates along with the 4 used in the model above. We saw that the regression coefficients obtained from all models were highly correlated (Figure 3.2). We decided to use the model with no principal components as covariates since it showed the highest number of enriched pathways among age-associated genes (Figure 3.3).

3.2.4 Evaluating empirical significance by permutation analysis

To estimate how many of the top age-related genes could be false positives, we permuted sample ages 100 times, and estimated the number of genes with P-value <

0.0001 in each permutation.

3.2.5 Pathway analysis

The set of 46,397 probesets was ranked by P-value and submitted to *LRpath* [123] for a directional enrichment analysis for pathways in GOCC, GOMF, GOBP and KEGG databases. *LRPath* allows users to submit genes ranked by their P-value (genes with multiple P-values are assigned the average P-value for that gene), and allows running both unidirectional (in which the direction of change of genes is not taken into account) and bidirectional tests (in which enrichment is tested separately in the subsets of up-regulated and down-regulated genes). Pathways with an FDR < 0.05 were considered significant in our analyses.

3.2.6 Constructing an age predictor for TM

To test if tissue age could be predicted using gene expression as a biomarker, we used 80-20 cross-validation. For 100 permutations, we generated a training dataset of 73 samples, and a test set of the remaining 20 samples. Then, we used the training set to identify genes associated with age: we identified these genes by building a linear model regressing expression of each gene to tissue age, and selecting those genes that showed a significant regression coefficient for age (significance was defined using different P-value thresholds ranging from 0.05 to 0.0001). Then, we used this list of genes to build a predictive model on the training set: predicting tissue age (the dependent variable) using expressions of the selected genes. This model was then used to predict tissue age on the test set.

The average predicted age for each sample across 100 permutations was taken as the ‘predicted age’. This was then compared to the actual sample age to get an estimate of model accuracy.

The above procedure was also used to build models to predict non-linear functions of age, such as age-square and log-age. We found that the linear models performed best, and do not show the results of the other two models. We also ran the same 80-20 cross-validation procedure on the 72 CSB samples, and found that the prediction performance decreased to a Spearman correlation coefficient of 0.38.

3.2.7 External databases and gene lists

A list of human aging-related genes [32] was obtained from [56] downloaded on August 30 2016. Genes which showed aging-related methylation signatures were obtained from the additional file 3 of Horvath, 2013. We obtained aging data from expression profiling in mouse tissues from the AGEMAP database [147]. We obtained tissue-specific regression coefficients from [148] accessed on August 29, 2016. We obtained pathways regulated in the mouse eye from http://cmgm.stanford.edu/ kimlab/aging_mouse/Supplemental%20Table%204.xls. We identified human orthologs for mouse genes using ENSEMBL Biomart [130], and then compared our gene list with the lifted over results from AGEMAP.

3.2.8 Tissue specificity of gene expression

The Tiger Expression Database [86] (source: bioinfo.wilmer.jhu.edu/tiger/ accessed in September 2016) uses ESTs to estimate tissue-specificity of gene expression.

3.2.9 Data availability

All expression data, sample metadata, and code used in the described analysis is available at https://github.com/shwetaramdas/aging.

3.3 Results

In this study, we collected TM tissue samples from 93 post-mortem samples with a broad range of ages (13-88). We used microarray profiling on Affymetrix arrays to measure genome-wide expression levels in each of these samples. Sample metadata are shown in Table 3.1. To identify genes whose expression levels were associated with age, we used *limma* using 4 variables as covariates (variables sex, RNA Integrity Number (RIN), ethnicity of the sample, and plate ID.) We excluded 11 samples obtained from the cornea, since these samples had lower RINs (Figure 3.1), and could possibly introduce a new level of confounding by virtue of having a different tissue composition. Adding additional variables to this 5-variable model did not change the regression coefficients for a majority of genes.

We found 2,822 probesets (corresponding to 2,324 genes) with a P-value of less than 0.05. None of these probesets was significant at a conservative threshold of 0.05 corrected for multiple testing using the False Discovery Rate (FDR). In further analyses, we define three levels of statistical significance; a highly stringent threshold of FDR $<$ 0.1 (one L1 gene), a less stringent threshold of P-value $<$ 0.0001 (eight L2 probesets corresponding to 8 genes), and a relaxed threshold of P-value $<$ 0.001 (83 L3 probesets corresponding to 74 genes.) Ranked P-values and regression coefficients of all probesets are listed on the project Github page [114].

We permuted the ages of the samples 100 times, and performed a regression analysis for each permuted dataset to estimate the null distribution of regression coefficients. The qq-plot of the average ranked regression coefficients across 100 permutations versus the qq-plot of actual ranked regression coefficients shows an inflation of high regression coefficients in our real dataset (Figure 3.4).

To visualize the expression pattern of the top age-associated genes, we generated a heatmap of the expression data (corrected for RIN, plate ID, sex and ethnicity) shown in Figure 3.5. We see a shift in gene expression patterns of these genes with age (samples are displayed in increasing order of age from bottom to top). In order to visualize the changes in gene expression of each gene, we present scatter plots of expression residuals with age for probes from the top 8 genes (Figure 3.6).

Previous studies on aging have shown that different tissues show different ratios of up-regulated to down-regulated genes with age [145]. The ratio of up-regulated to down-regulated genes in our dataset (51.7%) most closely resembles the adipose tissue when compared to an expression study on aging in seven different human tissues [145]. This number also supports the proportion of up-regulated genes found in the previous study on the TM [23].

3.3.1 Genes relevant to aging of the eye

The top 8 genes from our analysis (passing the L2 threshold) include those with functions relevant to the biological underpinnings of aging, described below. *Spopl* (Speckle Type BTB/POZ Protein Like) is part of a ubiquitin-ligase. These are protein complexes that target proteins for destruction by the proteasome. We see a decline in expression of *Spopl* with age, indicating a possible increase in the accumulation of damaged or misfolded proteins in the TM with age. The Affymetrix U219 array has four probes aligning to this gene, all of which have a similar range of regression coefficients, though the P-values for the other probes are lower (5.47e-0.3 to 9.19e-03).

Dnajc4 (DnaJ heat shock protein family (Hsp40) member C4) is a heat shock protein showing increasing expression with age. Hsp40 is a molecular chaperone whose function is to prevent proteins from misfolding. Its increased expression with age could represent a response to oxidative stress accumulation with age [22].

The *Prkcsh* (protein kinase C substrate 80K-H) gene is a glycan-processing enzyme. Glycoproteins form a major component of the cornea and the extracellular matrix of the TM, and previous studies have shown an increase in sulfated proteoglycans with age [24] leading to increased stiffness of the TM. The increased expression of *Prkcsh* could lead to this phenotype.

3.3.2 Pathway enrichment

We submitted the ranked list of genes (ranked by P-value) to *LRpath* to identify pathways that were enriched in aging (pathway list with P-values is on our Github page). Of the 520 pathways that showed enrichment, 304 were enriched among our set of up-regulated genes. LRpath does not correct for correlation structure between genes, which can lead to inflated test results. Thus, to get an estimate of the null distribution of P-values, we permuted sample ages 20 times, and submitted the ranked list of genes for each permutation to LRpath for pathway enrichment (we restricted the number of permutations to 20, since LRpath uses a visual interface requiring manual input of gene lists for each permutation). We found that the P-values of the top 9 pathways were not reached in any of the 20 permutations. We also found that for each of the top 520 pathways, the FDR was not reached for a similarly ranked pathway in any of the 20 permutations.

We see pathways relating to intracellular function are enriched among our set of down-regulated genes, and pathways relating to the extracellular part are enriched among our set of up-regulated genes (Figure 3). We also see pathways such as the electron transport chain, mitochondria, immune pathways, and the ribosome enriched in our differentially expressed genes.

3.3.3 Prediction of age

Previous studies have attempted to use gene expression as a biomarker of tissue age [48]. To test if gene expression in the TM could be used to predict age, we created 100 permutations of training and test sets of 75 and 18 samples respectively, and predicted sample ages based on the average prediction across all permutations. We built different prediction models based on linear, logged and squared values of age, to reflect different trajectories of expression changes with age; the linear models performed best. Our predicted values had a Pearson's correlation of 0.68 with actual ages

(Figure 7). When restricted to the 72 corneoscleral button (CSB) samples, our predictor had a spearman correlation coefficient of 0.38. Our predictor did not perform as well as other aging predictors built from DNA methylation profiles, but performed comparably with other methods used on gene expression data. Our relatively small size compared to other aging datasets could limit our predictive power.

3.4 Discussion

We find 8 genes whose expressions in the TM significantly change with age. We observe that the immune system and the extracellular matrix are most enriched among our age-associated genes. These pathways have been previously associated with aging signatures in other tissues, and are coherent with the physiology of the TM in particular. The TM is known to show increasing stiffness with age, and changes in both these pathways could lead to these changes. Increased oxidative stress in the TM has also been shown to lead to changed adhesion between cells in the TM and the extracellular matrix.

Amongst our top pathways were pathways that have been previously shown to be associated with aging in other tissues, such as the electron transport chain, mitochondria, immune pathways, and the ribosome. This shows the commonality of aging pathways across tissues despite there being no significant overlap between aging genes. However, while ribosomal genes show up-regulation with age in our study, they have been consistently shown to be down-regulated with age in the muscle and in other studies [12, 135]. We also see a host of ubiquitin-related pathways being down-regulated with age; this is concordant with the hypothesis that aging is correlated with an increase in misfolded proteins [13].

GenAge is a manually curated database of human aging genes, including the genes directly related to aging in humans and the best candidate genes obtained from model organisms [134]. One of our genes passing our L1 and L2 filters overlapped with the list of 305 GenAge human agerelated genes (TNFAIP8L1); an additional gene passing our L3 filter (F8A1) overlapped with the GenAge list; this is not a significant overlap using the hypergeometric test. Horvath [65] showed that the methylation signatures at 353 genes could be used to predict biological age. Of our top 353 genes, 6 genes were found in this list; this overlap is not significant. This suggests that most of our age-related genes do not harbor age-associated CpG methylation sites, and their association with age is regulated by mechanisms other than DNA methylation.

The AGEMAP resource [147] has a database of age-related genes and biological pathways in mice, compiled from gene expression studies on 8,932 genes in 16 different mice tissues. The regression coefficients for age from our analysis showed no correlation with the regression coefficients from expression analysis in any tissue in mice (for this comparison, we obtained the list of regression coefficients for mice). Previous analyses on the GTEx expression dataset across 9 human tissues have also shown that aging processes are not highly conserved between mice and humans [145] and the lack of correlation in our dataset supports that conclusion. However, we found all the top pathways enriched in the eye of mice in our analysis, except for heparin binding and lipid metabolism, suggesting a convergence in aging signatures at the pathway level.

There have been previous studies studying transcriptional changes with aging in other human tissues [78, 48, 149, 121]. When we compare the number of significant age-associated genes in our study with these previous studies, we find that all other tissues (kidney, muscle, skin, adipose and brain) except LCLs show much higher numbers of significant genes at the same significance threshold compared to our dataset. Specifically, when we look at a previous study on the aging human muscle with a sample size similar to our dataset, the range of effect sizes seen in muscle tissue is much higher than the range seen in our dataset: our top genes have a much smaller dependence on age in the TM compared to the muscle. Different tissues age at different rates; our results, along with the noisy association of our top age-associated genes with age, suggest that the aging signature is weak in the trabecular meshwork. This tissue is primarily composed of extracellular proteins. It is hence possible that age-related molecular changes are reflected more at the translational level than at the transcriptional level.

We do notice a trend that is opposite to that seen in previous studies on aging in other tissues. Previous studies have consistently shown a decrease in the expression of ribosomal genes with age. However, we see a significant enrichment of ribosomal pathways in our up-regulated genes. Some researchers have proposed an increase in ribosomal gene expression as a compensatory mechanism for decreasing efficiency with age [91]; this has also been seen as an increase in expression of ribosomal pathways in the brain with age [78]. It is possible that this compensatory mechanism is a cause for increased ribosomal expression in the TM.

The aging TM is known to have decreasing cellularity, which in turn leads to mechanical stretching of the TM [4]. Genes related to the extracellular matrix have been shown to be elevated in response to mechanical stretching in order to reduce resistance to aqueous humor outflow. The age-related increase in pathways of the extracellular space could point to this homeostatic response undertaken in response to mechanical stretching.

Figures

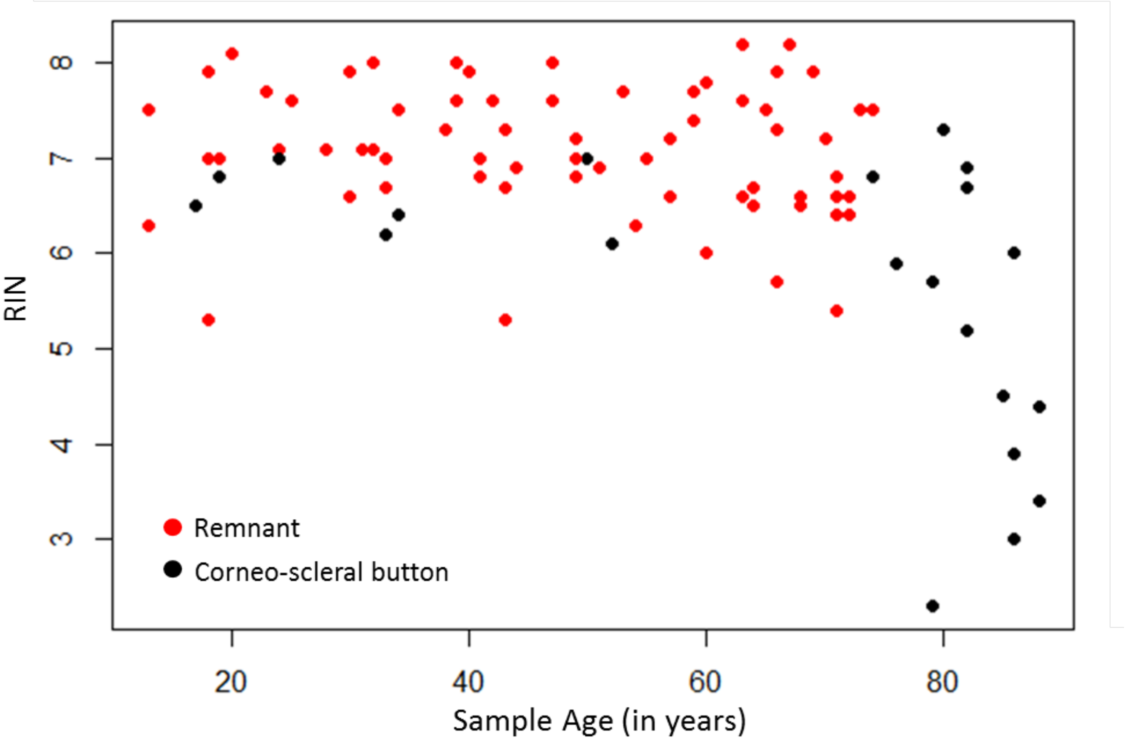

Figure 3.1: **Sample age versus RNA Integrity Number (RIN) for 93 samples** We see that the oldest samples appear to be the most degraded (low values of RIN)

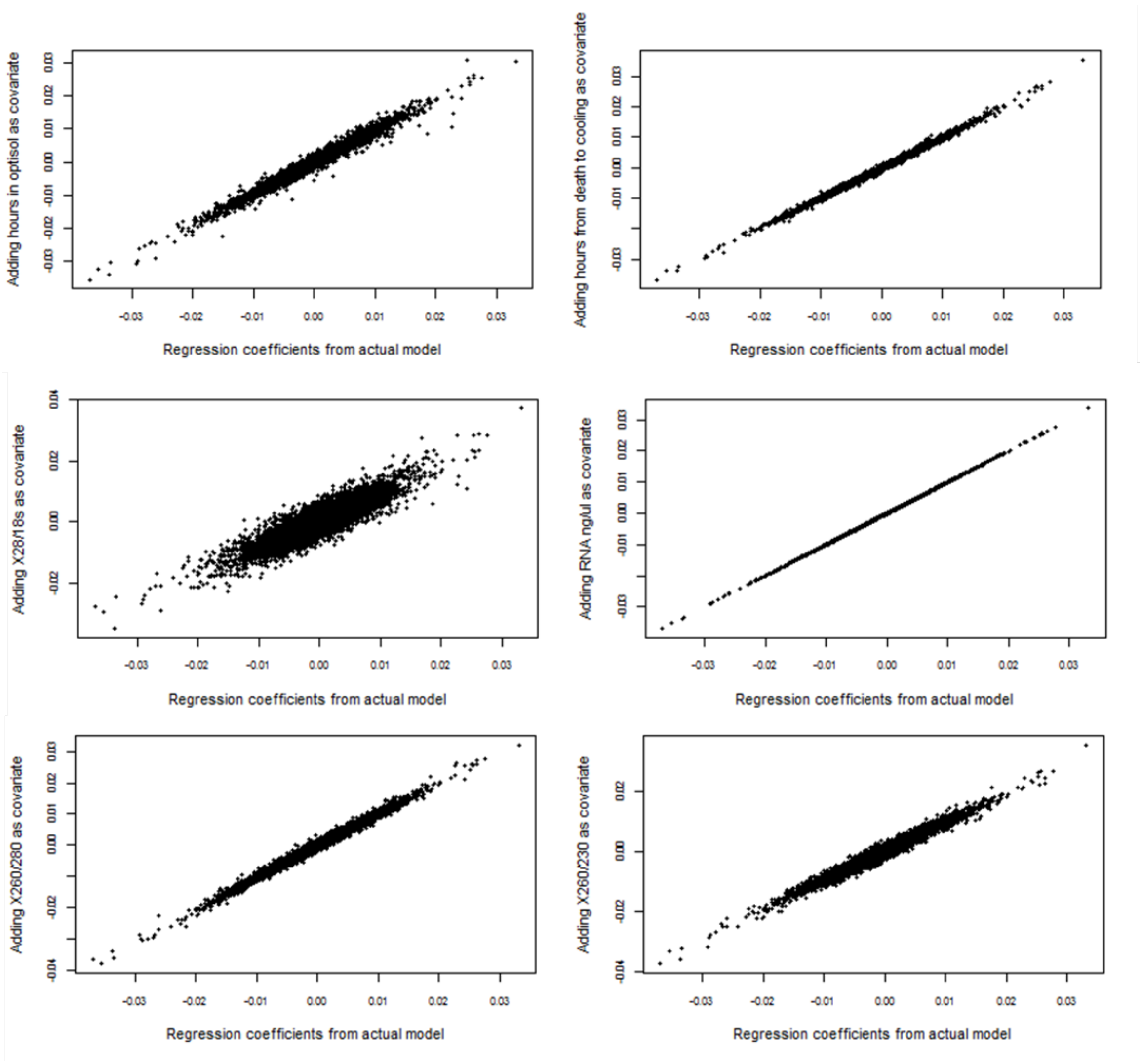

Figure 3.2: **Comparising regression coefficients between models** Shown above are the regression coefficients from our actual model (x axis) against those obtained from a new model incorporating an additional covariate (y axis). We see that adding additional covariates does not change the regression coefficients greatly.

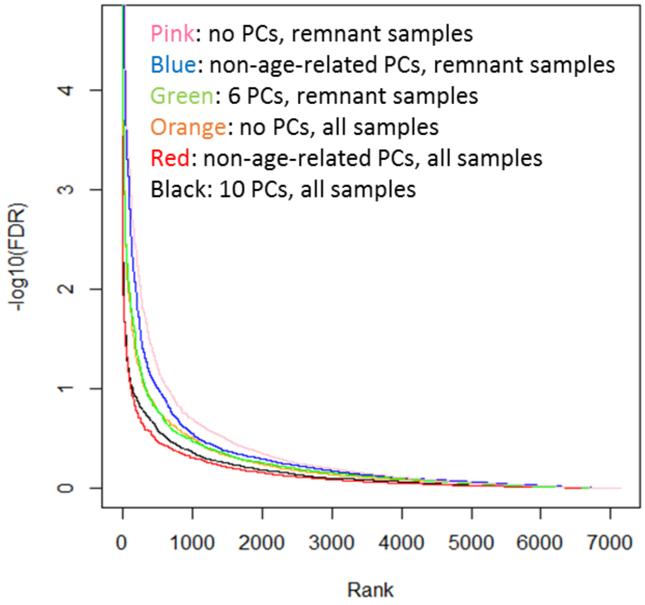

Figure 3.3: **Model with 0 PCs shows greatest pathway enrichment** Here we plot the FDR distribution from pathways enrichment run on the differential expression results of each model (with differing number of covariates used for correction in each model. We see that the pink line (representing 0 PCs used for correction) shows the highest number of significant pathways.)

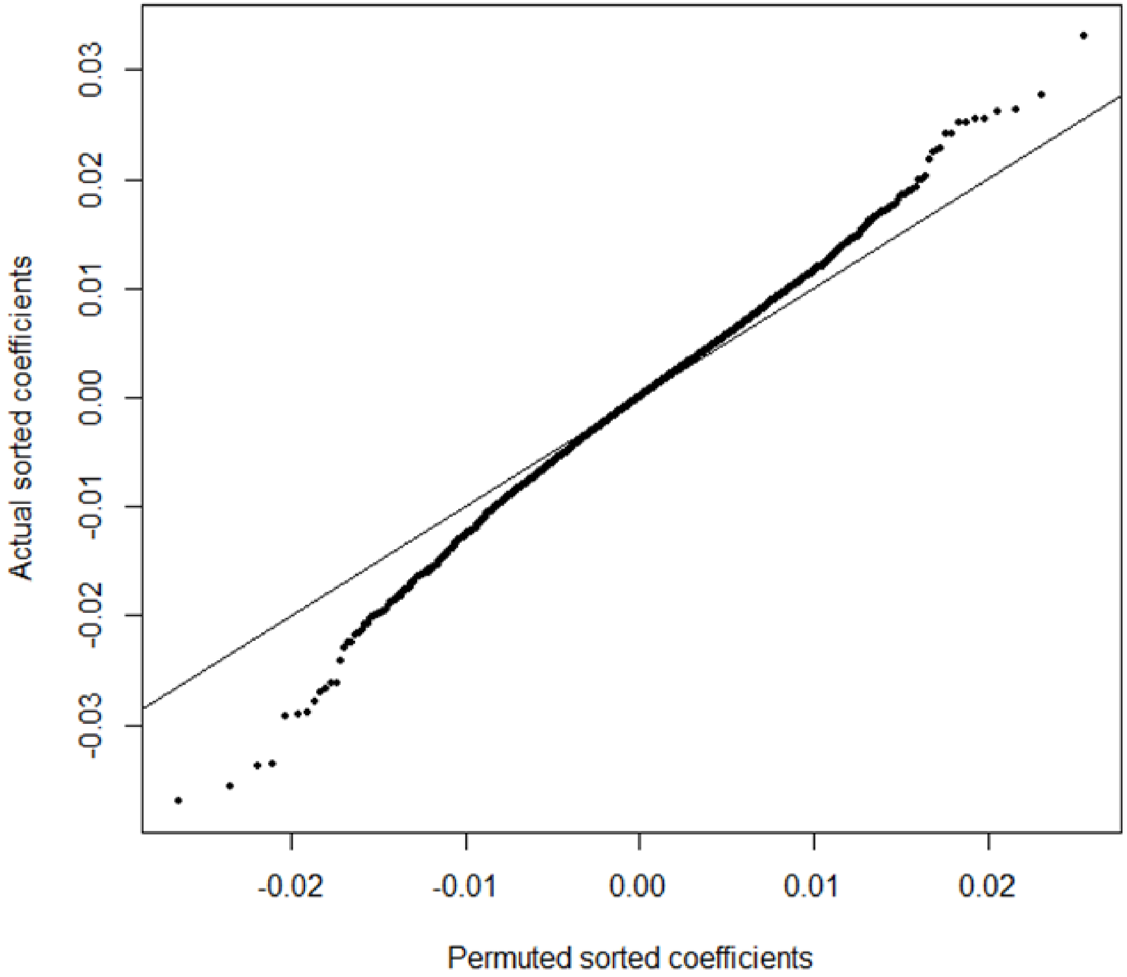

Figure 3.4: **Actual versus permuted regression coefficients** Plotting the true regression coefficents (y axis) against the mean ranked coefficients from 100 permutations.

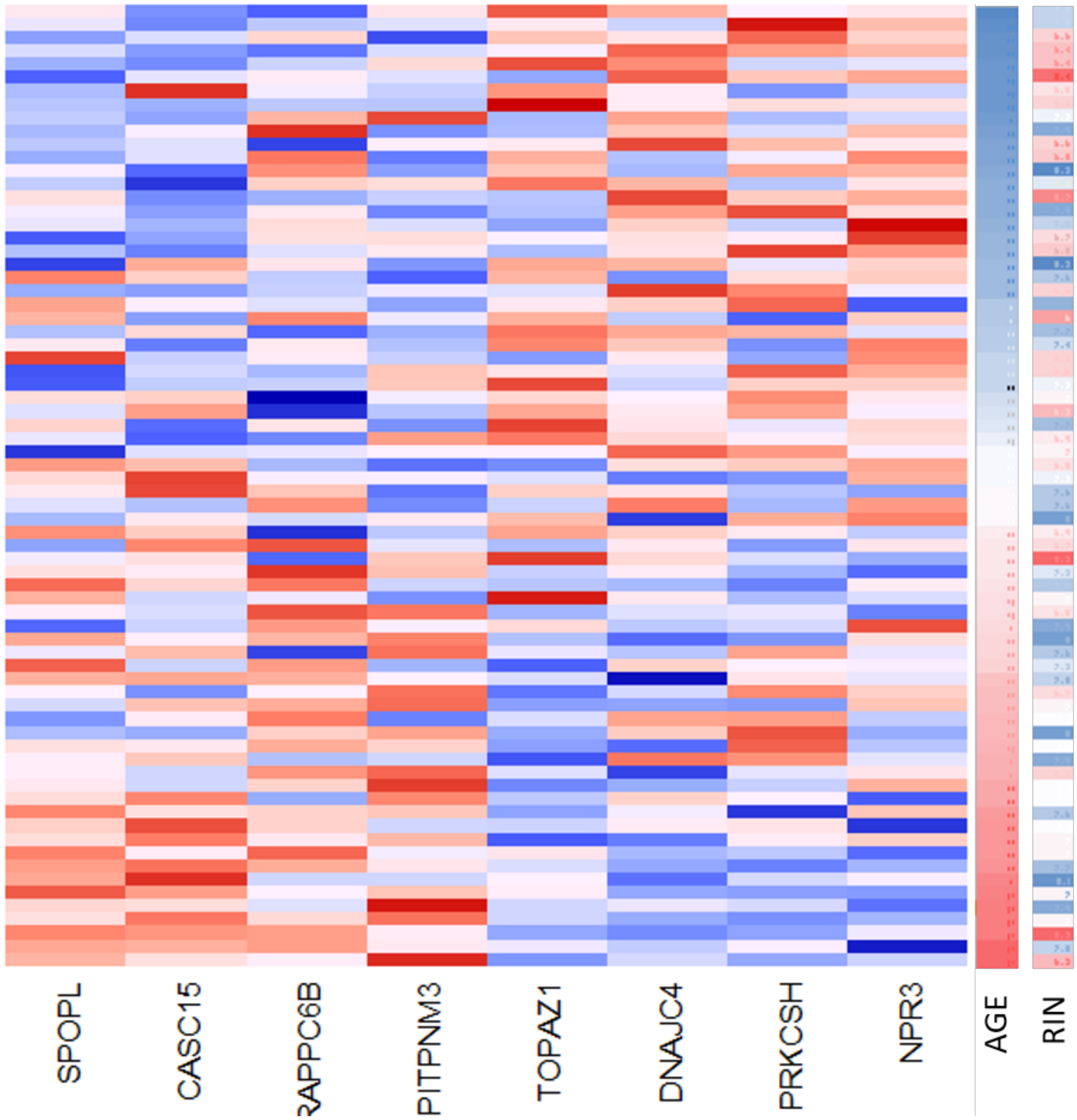

Figure 3.5: **Heatmap of standardized expression levels of top 8 probesets** We see that the top 8 probesets can be used to distinguish between young and old samples (panel on right shows age)

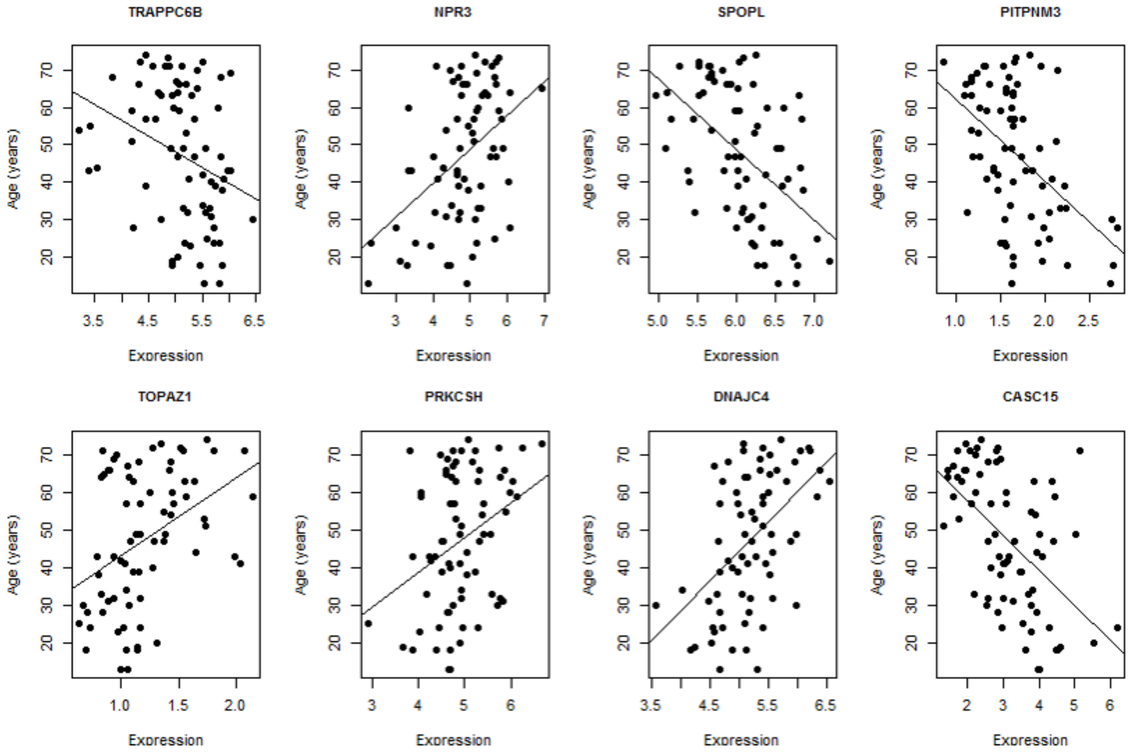

Figure 3.6: **Expression levels of the top 8 genes against sample age**

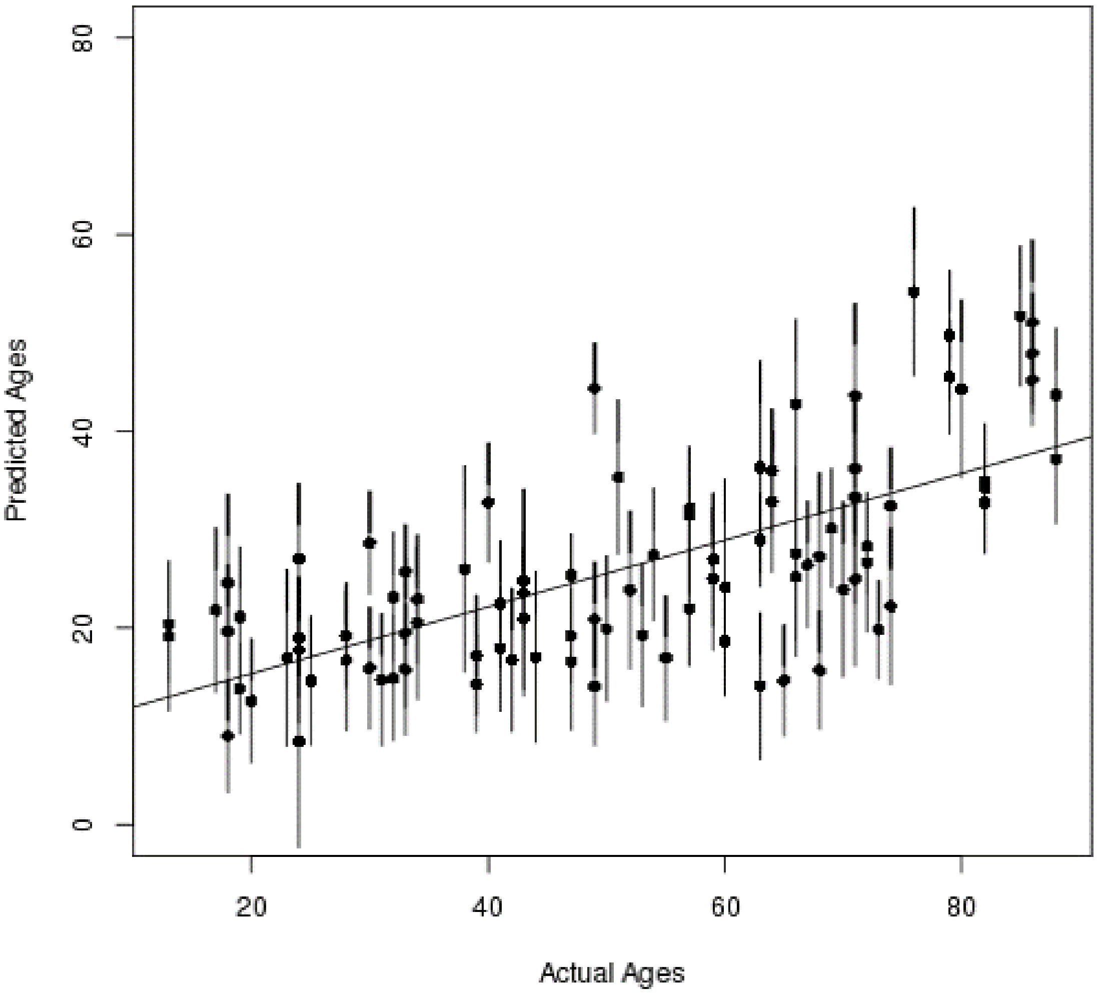

Figure 3.7: **Building a predictor for age** Our linear predictor for gene age shows a Spearman correlation of 0.68 (using all 93 samples,) and 0.38 (using only 72 samples from the CSB)

Tables

Quantitative				Categorical	
Variable	**Min**	**Median**	**Max**	**Variable**	**Categories**
Age	13	52	88	**Tissue Type**	Full cornea (21), Remnant (72)
RIN	2.3	7	8.2	**Sex**	Female (32), Male (61)
Hours in Optisol	6.6	97.5	223.2	**Race**	Caucasian (79), African-American (12), Asian (1), Indian (1)
x260/280	1.21	1.69	2.15	**Plate**	P1 (24), P2 (24), P3 (20), P4 (7), P5 (18)
Death to Extraction (hrs)	1.7	10.53	23.37		
Death to Cooling (hrs)	0	2.36	22.18		
Death to Dissection (hrs)	14	108	229		
ng/ul	4	11.1	23.37		
A260	0.099	0.277	3.868		
A280	0.075	0.165	2.484		
x260/280	1.21	1.69	2.15		
x260.230	0.06	0.57	2.06		

Table 3.1: **Sample properties**

CHAPTER IV

A Rat Model for Aerobic Capacity

4.1 Introduction

Aerobic capacity is the quintessential complex trait. Defined as the ability of the heart and lungs to get oxygen to the muscles, it has a heritability between 39% and 77% in humans [3], and is a phenotype with strong environmental influences. This phenotype is an excellent predictor of disease risk in humans, and peak exercise capacity is a better predictor of mortality than other established factors like smoking and diabetes[83].

In human studies of aerobic capacity, it is difficult to separate out the genetic and environmental contributions to the phenotype; i.e., separating the innate component from the trained component. It is not possible to obtain a cohort of samples completely controlled for any possibly confounding environmental factors such as diet and exercise. This makes genetic analysis particularly challenging because these environmental factors have long-term effects on aerobic capacity.

To overcome these challenges, researchers since 1996 [74] have established a rat model to study aerobic capacity and related traits. They started from a population of genetically heterogeneous rats derived from outcrossing eight inbred strains [58]. Following bidirectional divergent selection for untrained exercise capacity on this population, two rat lines were established—the high capacity runners (HCRs) and the low capacity runners (LCRs). These two lines differed not only in aerobic capacity but in a host of health-related traits, like blood pressure, body weight, oxidative stress, lifespan [118].

Despite extensive studies, the underlying molecular basis for the enhanced health of the HCRs and increased disease risk of LCRs eludes us. The LCRs and the HCRs have diverged significantly due to both selection and drift, and the effects of these two are difficult to separate out. In order to identify the true functional variants that are the actual targets of selection, we generated an F2 intercross of 650 samples, so that the neutral variants would no longer be associated with the selected phenotype.

4.2 Methods

4.2.1 Animals

The F2 intercross was performed in two batches. In the first batch, 4 males and 4 females were selected randomly from different families in generation 26 of each line, to generate 79 F1 rats from 8 HCR-LCR pairs. 20 males and 20 females were randomly picked form these F1s to form mating pairs, generating 154 F2 rats. In the second batch, 9 males and 9 females were selected from different families in generation 28 of each line to form 18 mating pairs, which generated 163 F1s. From these animals, 43 males and 43 females were selected and mated to yield 491 F2 rats.

4.2.2 Treadmill yest to measure AEC

Eleven week old animals were subjected to run-to-exhaustion tests without prior training, except for brief sessions of treadmill education during the week prior to the tests. For the run-to-exhaustion test, each trial starts at a velocity of 10miles/minute, which increases by 1 minute/mile every 2 minutes until the rat reaches exhaustion. Exhaustion point is defined as the third time a rat can no longer keep pace with the treadmill and remains on the shock grid for two seconds rather than resuming running. The phenotype distribution for best running distance is shown in Figure 5.1.

4.2.3 Genotyping

A custom Affymetrix Axiom panel of ˜700,000 SNPs was designed to genotype our F2 animals; SNPs in this panel were selected based on criteria previously described in Ren et al, 2016. These SNPs were based on the rn5 version of the rat genome assembly. After lifting over to the rn6 genome, we were left with 380,990 autosomal SNPs (none of the SNPs mapped to chromosome X).

4.2.4 RNA-Sequencing and analysis of associated genes

We generated RNA-Seq data from the skeletal muscles of 434 F2 animals at rest. Reads were aligned to the rn6 genome using *STAR* version 2.5.2a [35], and read counts obtained using *HTSeq* version 0.6.1p1 [6]. We obtained read counts for 32,734 genes. 8,760 genes had read counts of 0 in all samples and were removed (23,994 genes with counts >0 in at least 1 sample). We use log of the counts per million (CPM) (with 1 added to deal with zero counts) as the metric of gene expression. We show the first few principal components of gene expression in figure.

Xist expression was used to confirm sample sexes for quality checking. Two animals were found to be incorrectly labeled, and were removed.

4.2.5 PEER Factors to correct expression data

Expression data can have hidden structural confounders leading to unwanted sources of variation. One way to account for these confounders is to use principal component analysis, and correct out the top PCs. Stegle *et al* [133] have proposed a Bayesian approach to this correction using a method called PEER, which shows higher power than PCA-based correction. There is no standard approach to determine how many PEER factors are to be taken out. We used eight PEER factors because this number maximized the number of cis-eQTLs on chromosome 10.

4.2.6 QTL analysis

We used EMMAX [72] to run a linear mixed model on our phenotype data corrected for sex and batch. EMMAX corrects for genetic relatedness before calculating associations between genetic variants and phenotypes which is essential in such a structured dataset, which could otherwise show false associations due to family structure.

4.2.7 eQTL analysis

We ran eQTL analysis using Matrix-eQTL. As input to Matrix-eQTL, the phenotype was inverse-normalized counts per million (CPM). We regressed the inverse-normalized CPM against the genotype, using as residuals sex, family ID, and the top 8 PEER factors. We restricted our analysis to cis-eQTLs—eQTLs within 1Mb of the start and end co-ordinates of each gene.

4.3 Results

4.3.1 Distribution of phenotype in F2s

In Figure 4.1, we see the phenotypic distribution in different generations—F0, F1 and F2. As expected, we see a higher variance in phenotype in F2s compared to the F1; however, the increased mean in the F2s suggests a possible dominance component to the heritability. We also see a difference in phenotypes between the animals from the two batches (Figure 4.2), with animals from batch 2 showing higher values of AEC.

4.3.2 Genetic Structure in Our Dataset

Our F2 population is highly structured, with multiple levels of structure. The two batches are derived from two generations (26 and 28) of the HCR-LCR population, and thus show genetic differentiationthe first few principal components show a separation between batches 4.3. Another level of structure is the relationships between F2s due to multiple animals in the same litter. Accounting for this structure is important in downstream analysis (for which we use Family ID as covariates, or mixed models incorporating genetic covariance).

4.3.3 Heritability

We show the correlation between the average parental running distance and the child's running distance in Figure 4.4. We estimated phenotype heritability both using SOLAR [75], and using GCTA [146] for SNP-based heritability. The SNP-based heritability estimates are 0.54 (with SE of 0.06), and the pedigree-based estimates of heritability are 0.60 (with SE of 0.05). However, SNP-based heritability estimates are known to be affected by population stratification, and given the level of structure in our dataset, it is likely that this is an over-estimate of the true heritability.

4.3.4 QTL analysis

The results from the linear mixed model show a deflated QQ-plot 4.5, showing that we are under-powered to detect true signals. This is a negative result that nevertheless gives us a hint of the polygenicity of this trait in this population. This is contrary to the number of causal loci predicted by the Castle-Wright estimator (< 15), which makes strong assumptions about independent causal loci having the same effect sizes.

If we restricted QTL analysis to SNPs that were eQTLs (P-value < 0.001), the deflation remained the same. Because of the deflated signal of the QTL results, we looked at them in the light of other results on the same samples (described later.)

4.3.5 eQTL analysis

We ran Matrix-eQTL [126] to estimate main-effect cis-eQTLs for each of 16,520 genes in our dataset. Restricting SNPs to those within 1Mb of a gene, we found cis-eQTL-eGene pairs (defined by an FDR cutoff of 0.05). We will have to perform further conditional analysis to detect the number of independent eQTL signals per gene. 5,975 genes (36.2%) were eGenes (i.e., had at least one SNP that was an eQTL), and 265,582 SNPs are eQTLs (it is likely many of these are not causal but in linkage disequilibrium with the causal eQTL).

4.3.6 Gene expression analysis

Are there genes whose expression levels are significantly associated with AEC? We had 409 animals from batch 2 with phenotype information and RNA-Seq data. We used CPM (counts per million) and removed genes with non-zero counts in fewer than 300 samples, leaving us with 16,520 genes. For each gene, we performed a linear regression with AEC as the phenotype, using sex, family ID, two PEER factors (described in methods) un-correlated with the phenotype, and expression principal components not associated with the phenotype as covariates. The QQ-plot shows considerable inflation 4.6, but 14 genes passed an FDR threshold of 0.05 (Table 4.1).

We compared our list of regression coefficients with the HCR-LCR fold changes obtained from microarray analysis and differential expression analysis between 24 LCRs and 24 HCRs at rest (described [118]). We see a high correlation between the regression coefficients and the fold changes 4.8. This compelled us to run a meta-analysis between the two gene expression studies. We used a signed P-value-based meta-analysis [37] for the meta-analysis—using P-values from edge 8 of the model in [118], representing differential gene expression between young LCRs and HCRs at rest. This gave us 28 genes with an FDR < 0.01.

We performed directional pathway enrichment analysis of the differentially expressed genes using *LRPath* [123] by submitting a ranked list of genes, with P-values from the meta-analysis. We found 355 significant pathways (not independent) at an FDR threshold of 0.01. Genes related to fatty acid and lipid metabolism are most likely to be associated with the phenotype. Other implicated pathways include the genes in the mitochondrion, and the extracellular matrix, both of which are up-regulated with increasing AEC. Pathways related to the ribosome tend to be enriched in genes down-regulated with increasing AEC.

4.3.7 Integrating results

To prioritize our list of genes significantly associated with expression, we look at our top associated genes from the meta-analysis to identify eQTLs for these. Of the 28 significant genes (FDR <0.01), 12 have significant eQTLs, and 6 of these 12 genes have nominally significant eQTLs (P-value < 0.05). These are *Raver2, Acadsb, Vtcn1, Pcyt1a, Cox15, Rab11fip3.*

From this list of six genes, we first prioritize gene *Acadsb* because of its role in valine/isoleucine metabolism, which is the top pathway enriched in differential expression. *ACADSB* is an enzyme that processes branched-chain amino acids in the mitochondrion, particularly valine and isoleucine. Figure 4.9 shows the QTL and eQTL associations for a SNP at the 5' end of *Acadsb*.

We also have orthogonal evidence supporting the role of *Acadsb*—initial analysis of metabolite data from SILAC assays (unpublished data) shows us that isoleucine levels are significantly associated with best running distance in both sexes. Moreover, we see increased allele frequency differences between the LCRs and HCRs at this gene in generation 26 compared to generation 5 [117], indicating that this gene could be a target for selection.

4.4 Discussion and ongoing work

A philosophical question underlying this project is the selection itself—are we truly capturing the traits we intend to study? The treadmill test could be capturing phenotypes unrelated to intrinsic aerobic capacity, such as pain tolerance. Our results (and results from similar analyses in the LCR-HCR population) are one indication of our being on the right path, since the top pathways associated with our phenotype are metabolic in nature, and are relevant to the physiology of the muscle.

Orthogonal studies on the LCRs and HCRs have shown that these two lines metabolize fat differently—this is further borne out by our initial results in our F2 population—with fatty acid metabolic pathways and genes having the strongest signal.

This is a rich dataset, which allows us the opportunities to fine-map causal variants using integrative approaches. We present here a first-analysis on this dataset. We have identified potential candidate genes to test *in vitro* and in animal models. Ongoing and future work will focus on incorporating and mining all possible sources of complementary information, including whole-genome sequencing data from a small subset of HCRs and LCRs, whole genome sequencing data on the eight founder strains, and looking more deeply into regions of the genome what show signatures of selection between the two lines. We will also look at the other gene candidates (some of which are mitochondrial, like the *Cox15* gene, and ribosomal, like the *Raver2* gene), including those obtained using other significance thresholds. We are also currently obtaining ATAC-Seq data on a subset of these F2 animals, allowing us to dig into the non-coding space for influences of non-coding regulation on AEC.

Besides being a useful model for fine-mapping variants for AEC, this dataset presents many opportunities to understand the basic biology of the skeletal muscle (this is the most highly-powered expression dataset from this tissue for rats) and perform comparative studies of muscle metabolism. We have already seen fundamental differences in muscle physiology between rats and humans—while the gene *Myh7* can be used to differentiate males and females in humans (FUSION consortium, unpublished data), we do not see this differentiation in rats in our dataset. Mining this dataset with related genomic, proteomic and metabolic resources will give us a deeper understanding of aerobic capacity, and muscle physiology.

Figures

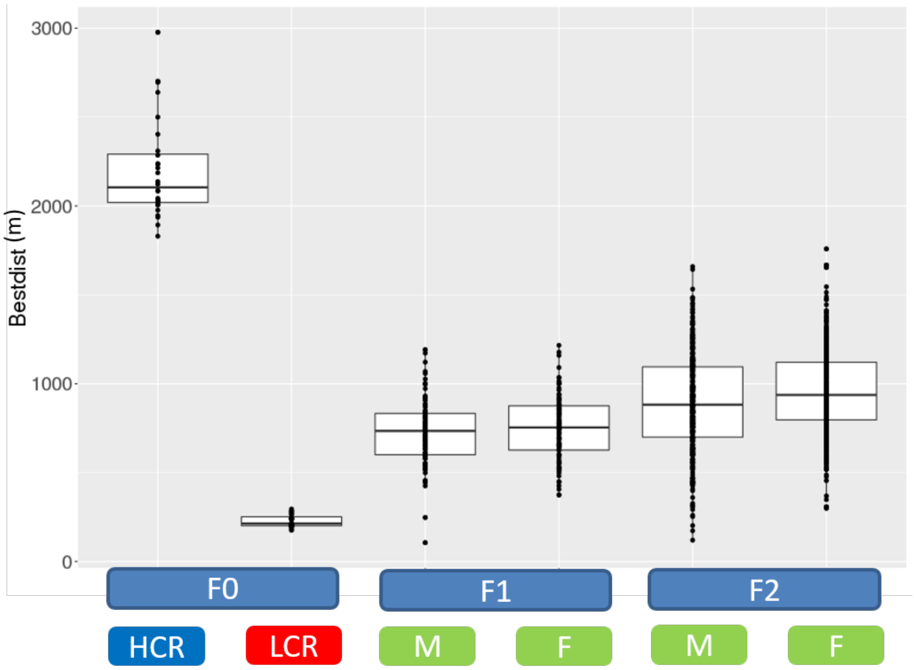

Figure 4.1: **Phenotype distribution in our F2 animals**

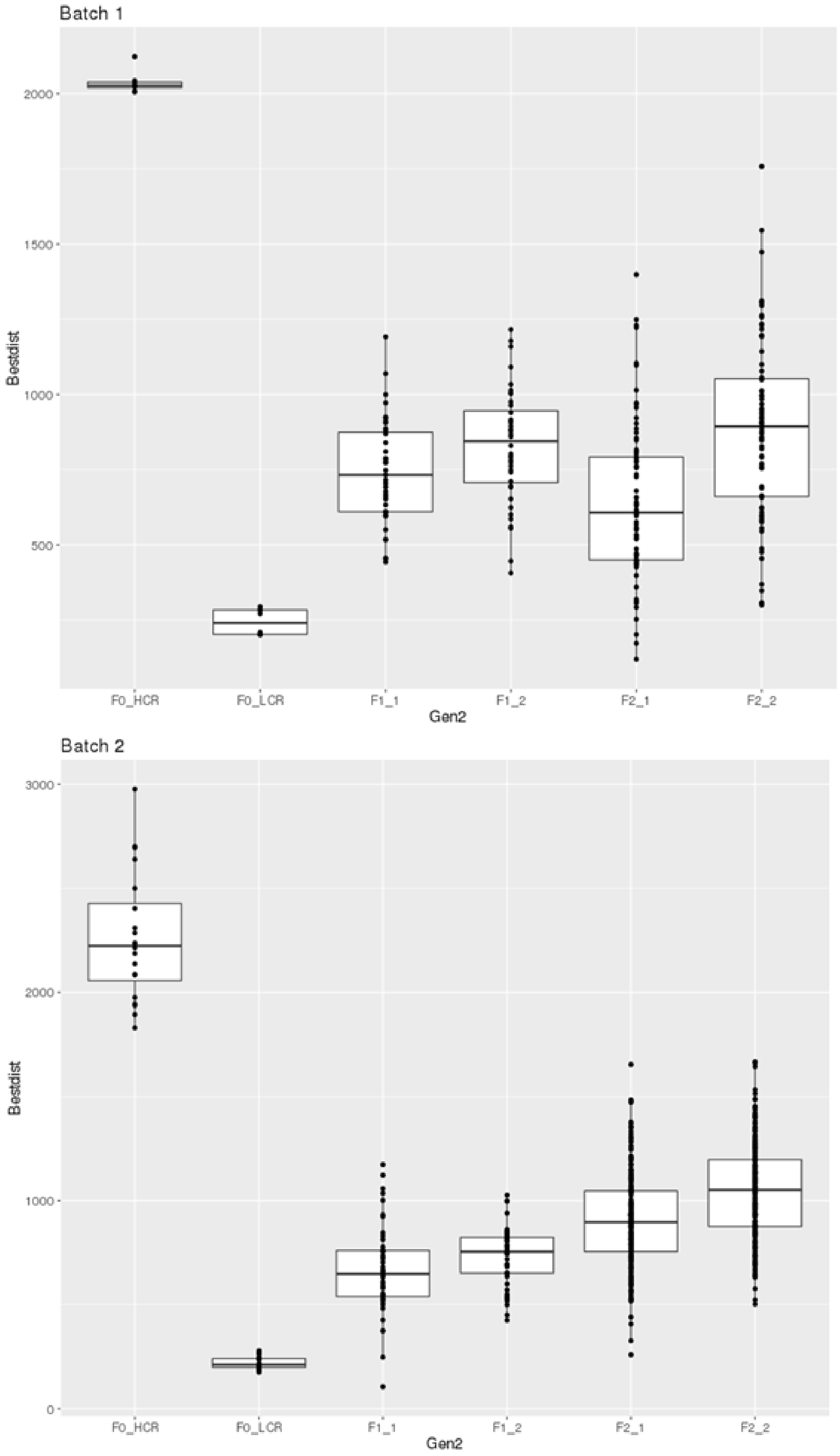

Figure 4.2: **Phenotype distribution for the two batches** The top panel shows phenotype distribution in batch 1, and the bottom panel shows phenotype distribution in batch 2. We see that the distribution of running distance in F2s from the second batch (bottom panel) has a higher median than that in F2s from Batch 1 (top panel).

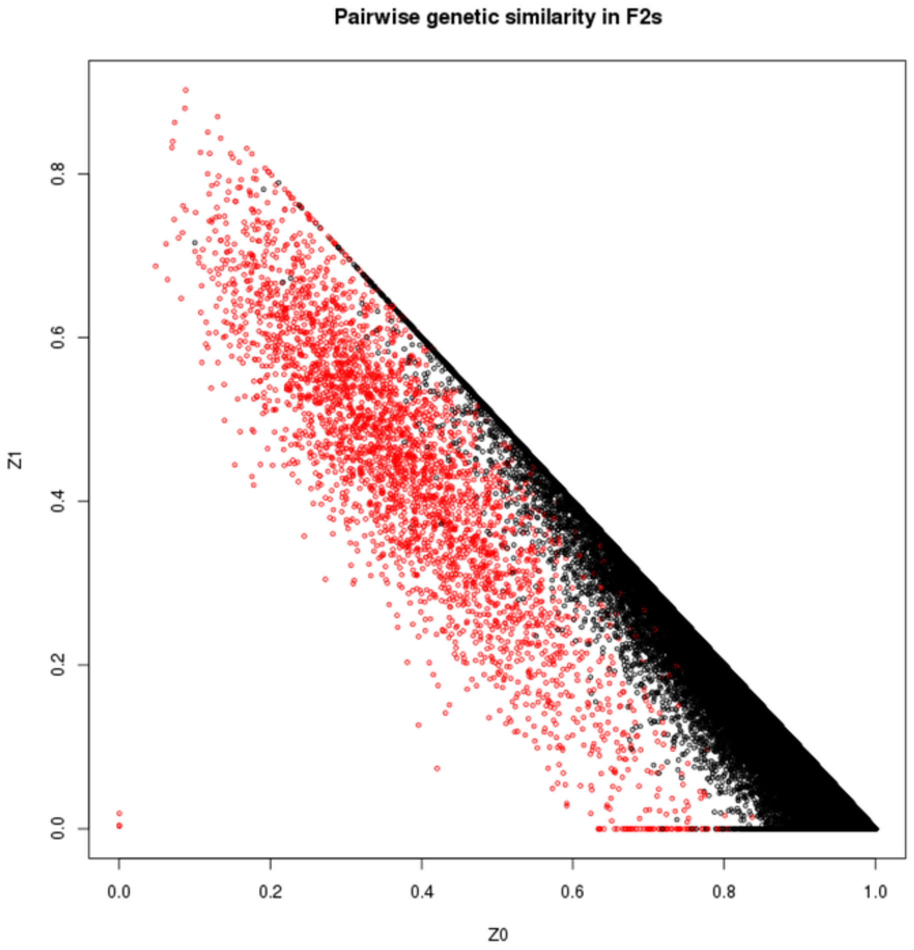

Figure 4.3: **Pairwise Genetic Similarity in F2s.** Z0 and Z1 are shown on the X and Y axis respectively, where Z0 is the probability of IBD 0 and Z1 is the probability of IBD 1. Siblings are colored in red, and 2-nd degree or more distant relatives are colored in black.

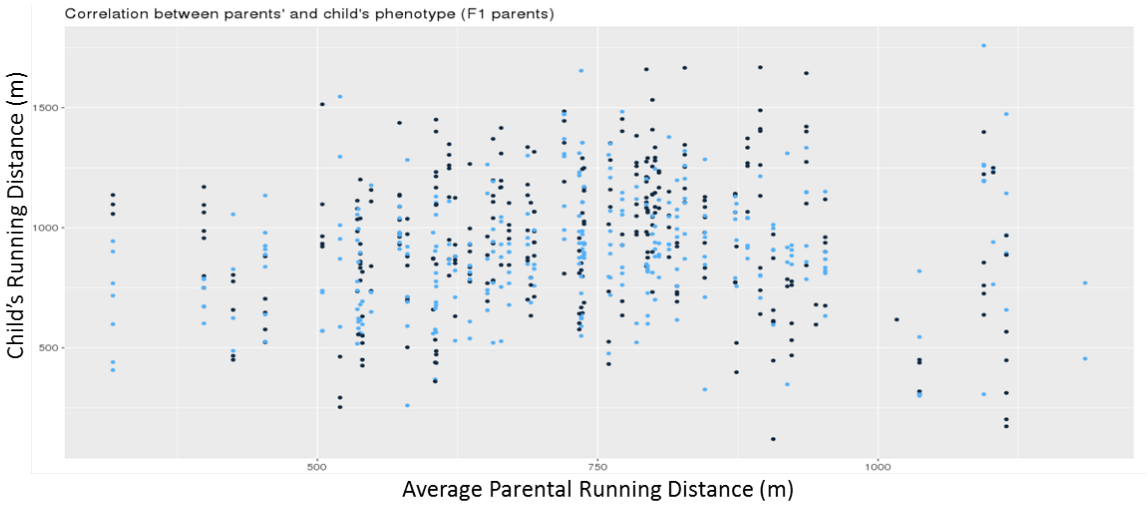

Figure 4.4: **Correlation between average parental phenotype and child's phenotype**. Female offspring (F2) are coloured in black, and males in blue. The average running distance of the two F1 parents is shown on the x axis, and the child (F2) phenotype on the Y axis.

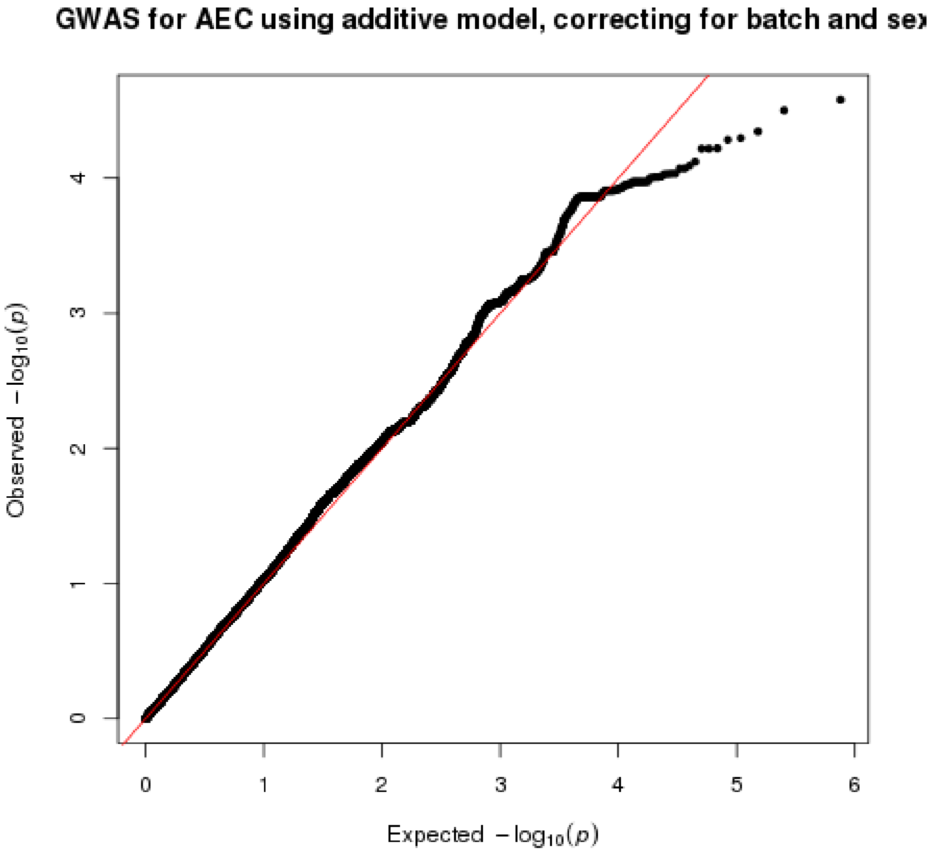

Figure 4.5: **QQ Plot**QQ Plot from a linear mixed model association analysis for AEC shows deflation.

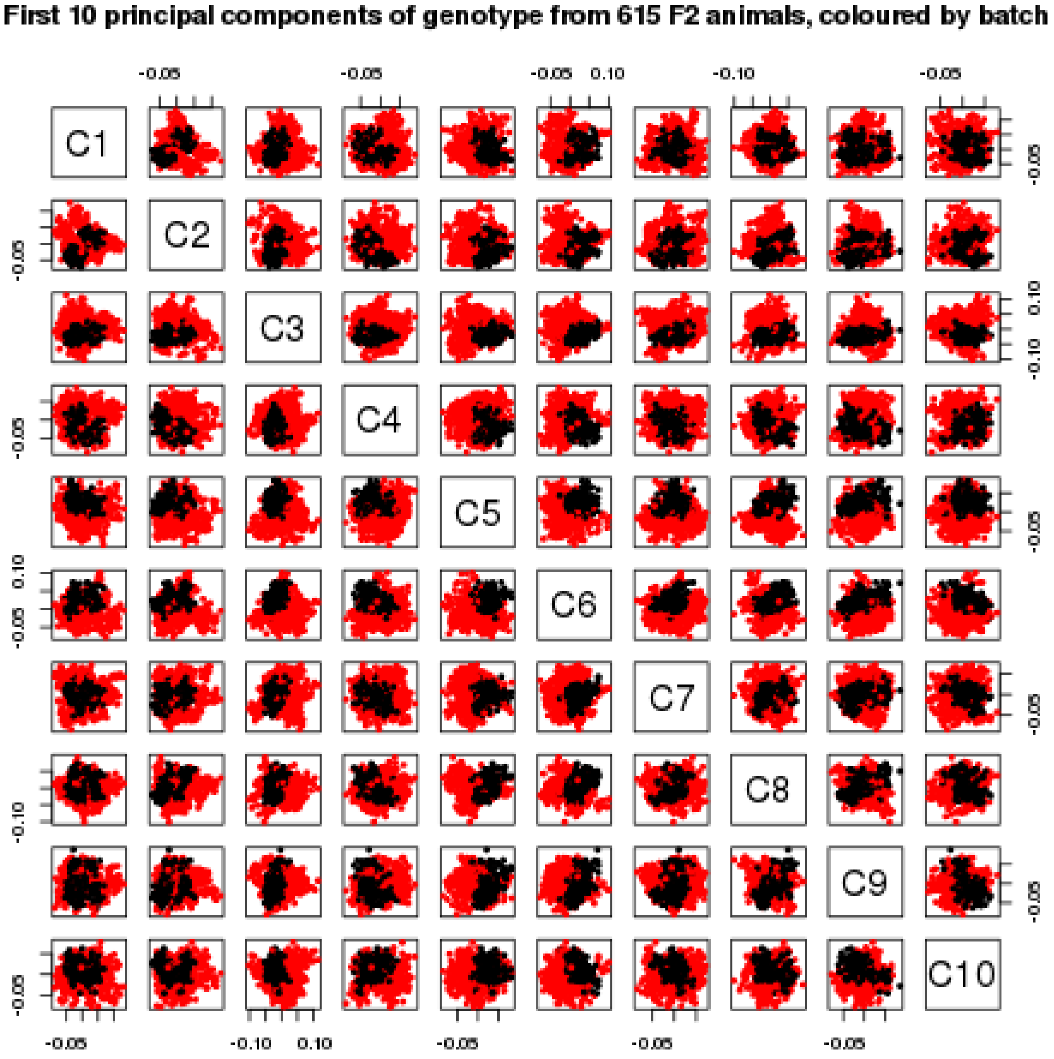

Figure 4.6: **MDS plot** Top 10 principal components of the genotyping data from our 615 F2 animals, colored by batch.

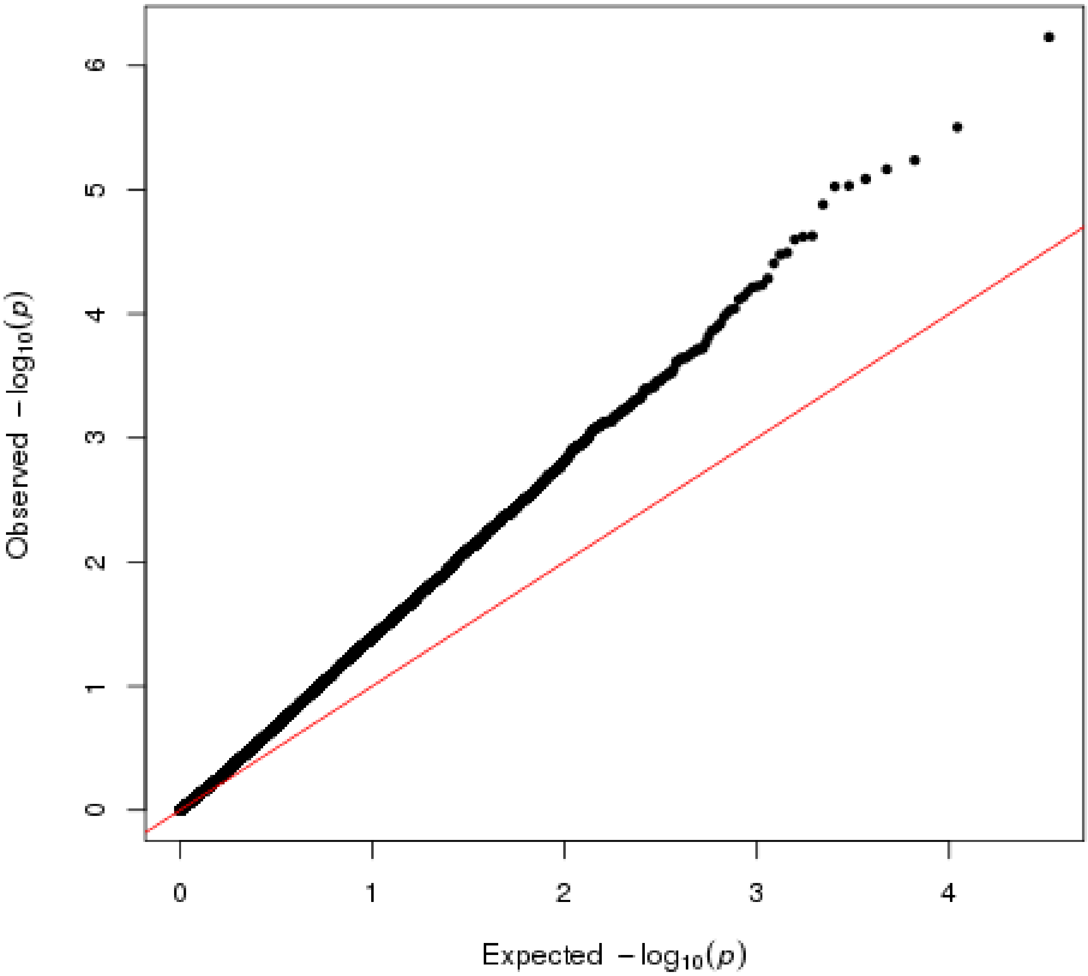

Figure 4.7: **QQ Plot for genes whose expressions are associated with AEC**

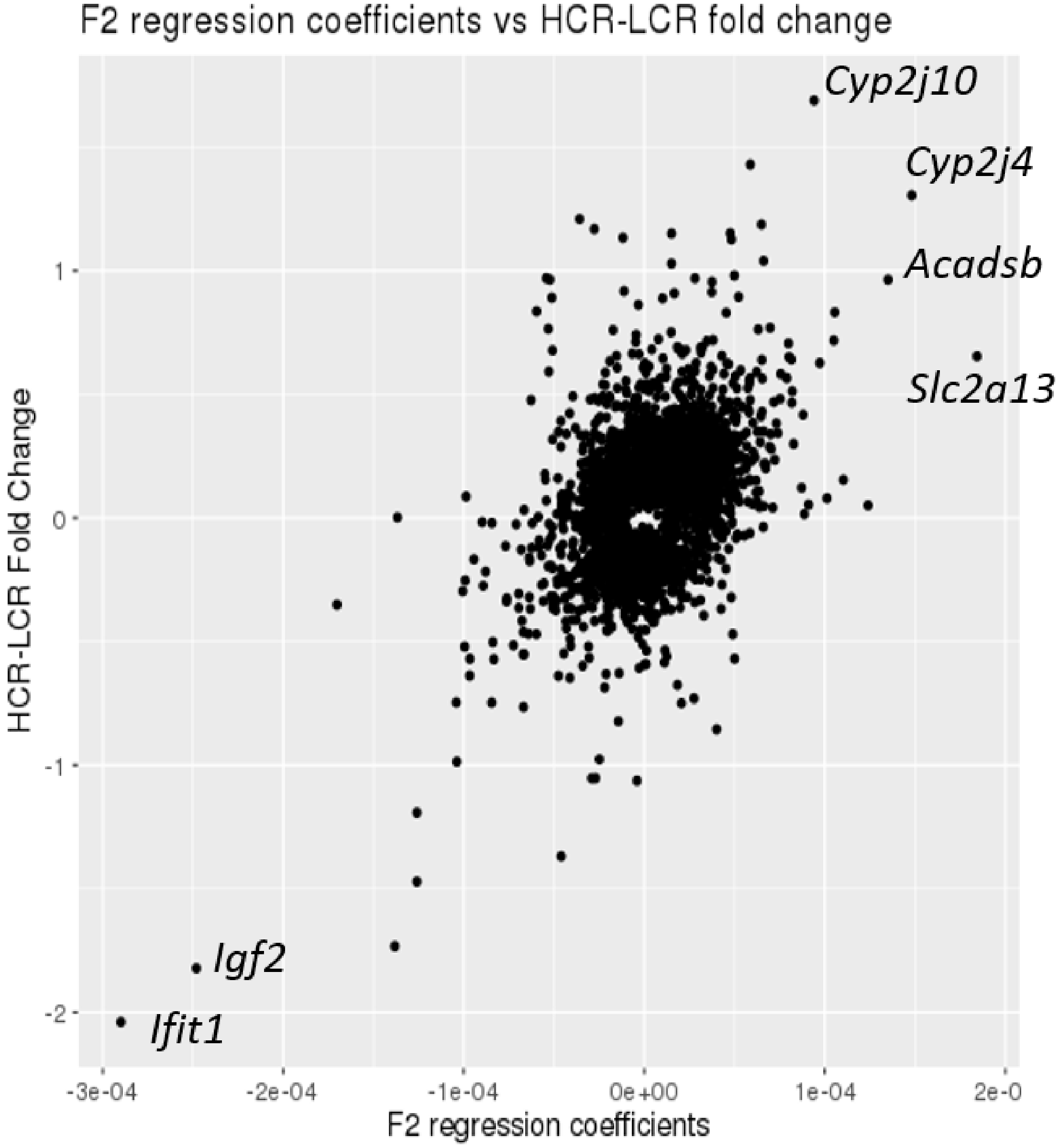

Figure 4.8: **Comparing regression coefficients in F2s with HCR-LCR fold change** We see a spearman correlation of 0.43. The top genes are labeled in the plot.

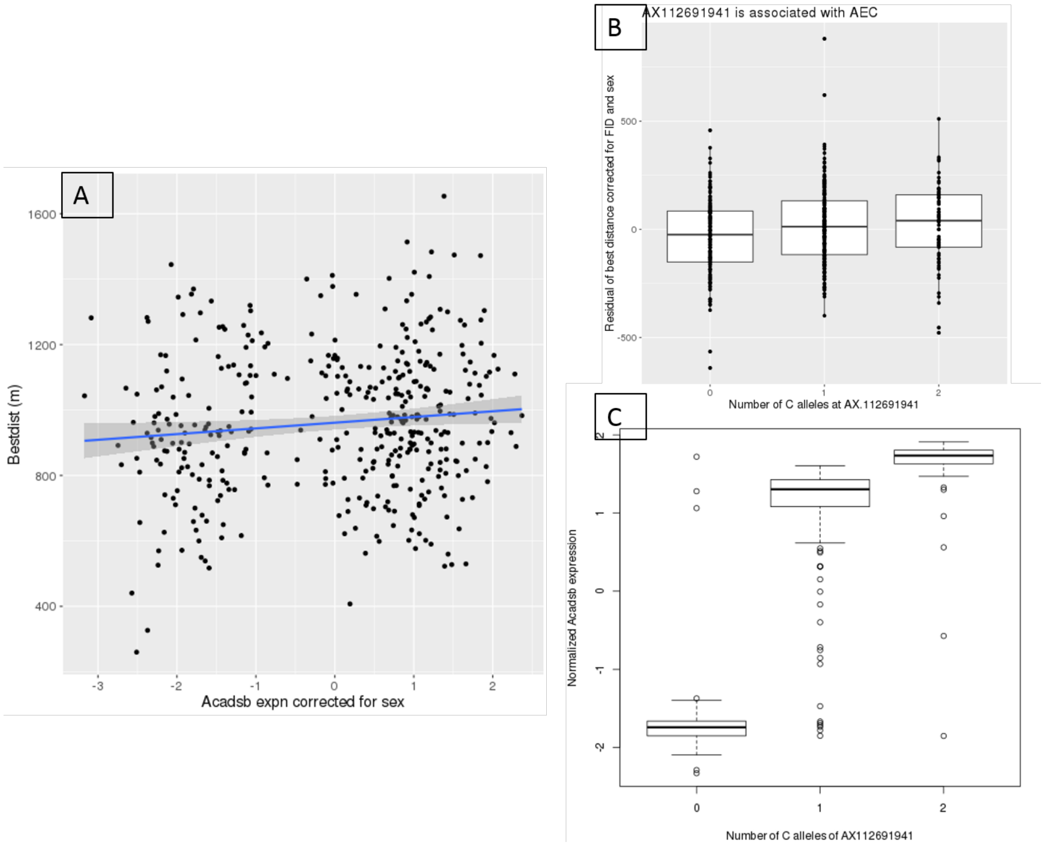

Figure 4.9: **Gene Acadsb is associated with AEC**. Panel A shows the association of normalized Acadsb expression (corrected for sex) with AEX. Panel B shows the QTL association for SNP AX-112691941 with AEC. Panel C shows the association of SNP AX-112691941 with normalized *Acadsb* expression.

Tables

Gene Symbol	**Gene Name**
Vps37b	Vacuolar Protein Sorting 37 Homolog B
Hspa2	Heat Shock Protein family A (Hsp70) Member 2
C1galt1	Core 1 Synthase, Glycoprotein-N-Acetylgalactosamine 3-Beta-Galactosyltransferase Hbp1 HMG-Box Transcription Factor 1
Naa15	N-alpha-acetyltransferase 15
Ghitm	Growth hormone-transmembrane protein
Vtcn1	V-Set Domain Containing T-Cell Activation Inhibitor 1
Gab1	GRB2-associated-binding protein 1
Ypel3	Yippee-like 3
Eif5b	Eukaryotic Initiation Factor 5b
Nifk	Nucleolar Protein Interacting With the FHA Domain of MKI67
A930018M24Rik	-
Gspt1	G1 to S phase transition 1
Dync1h1	Cytoplasmic dynein 1 heavy chain 1Imported

Table 4.1: **List of genes whose expressions are associated with AEC**

CHAPTER V

Extended Regions of Suspected Mis-assembly in the Rat Reference Genome

5.1 Background

A reference genome is the representative example of a species' genome sequence. *Saccharomyces cerevisiae* (baker's yeast) was the first eukaryote to have its genome sequenced. Since then, the genomic era has led to a host of genome projects to sequence the genomes of diverse organisms, including more than 800 bacteria and 100 eukaryotes. These reference genomes allow comparative genomics approaches, but also serve as a useful map or scaffold in assembling the genomes of individual samples within that species.

The laboratory rat (*Rattus norvegicus*) is an important model organism for studying the genetic and functional basis of physiological traits. Compared to the mouse, the rat shows a greater similarity to humans [68] and has been widely used in physiological, behavioral and pharmacological research. With the arrival of high-throughput genotyping and sequencing technologies, the rat has also been used in genetic studies to map causal loci, or identify genes that affect disease-related traits.

A fundamental resource in such genetic studies is the rat reference genome [46], which provides the coordinate system to organize our rapidly increasing knowledge of rat genes, their regulatory elements and functional profiles of diverse tissues, as well as gene variants and dysregulation in disease models. The reference genome is also the basic map in comparative analyses that focus on the evolutionary relationship among rat strains or between the rat and other organisms.

Gene discovery studies can be roughly classified by the type of mapping populations adopted. Currently, popular genetic systems include naturally occurring outbred populations, laboratory-maintained diversity outbred populations, inbred line-based crosses (e.g., F2-crosses, or advance inbred lines), recombinant inbred lines, and many others. Regardless of the system, a comprehensive knowledge of DNA variation in the mapping population is essential for both the study design and biological interpretation. In this study, we sought to use whole-genome sequencing (WGS) to ascertain DNA variants in eight inbred strains: ACI, BN, BUF, F344, M520, MR, WKY and WN, which are founders of the Heterogeneous Stock (HS) population. The HS rat has been used for genetic studies of metabolic and behavioral traits [132]. WGS data for these eight strains have been previously described [14, 61], using the SOLiD technology. Here we present WGS results from the Illumina technology, containing genotypes at more than 16.4 million single-nucleotide variant (SNV) sites. We expect that the sequences of the eight HS founders and fully-ascertained DNA variations can aid the imputation, haplotyping, and fine mapping efforts by the rat genomics community.

When analyzing the SNV data we noted that, while eight founders are inbred, all contain an unusually high amount of heterozygous nucleotide positions. Remarkably, these sites tend to concentrate in hundreds of discrete genomic regions, which collectively span 6-9% of the genome. We show that the heterozygous genotypes tend to recur in multiple, if not all, of the eight strains, and that the suspected regions tend to have higher-than-average read depths. We propose that these regions can be explained by mis-assembly of the rat reference genome, where many of the highly repetitive segments have been erroneously “folded” in the current coordinate system. This interpretation is not unexpected when one compares the genome assembly statistics between mouse and rat: the latest release of the mouse reference genome (GRCm38.p67) contains 885 contigs and a median contig length of 32.3 megabases; whereas the rat reference genome (rn6) has 75,687 contigs and an median length of 100.5 kilobases. With this report we release mask files for the suspected regions, so that they can be used to flag questionably results in current genomic studies until the time when a revised, more accurate reference assembly becomes available.

5.2 Methods

5.2.1 Animals, DNA samples, whole-genome sequencing

Eight animals, one for each of the eight founders of the HS population, were used in the study. Genomic DNA was extracted at the University of Michigan (Ann Arbor, MI) from the liver of seven animals (ACI/N, BUF/N, F344/N, M520/N, MR/N, WKY/N, WN/N), and from the tail of a BN/N animal. All animals were female. The DNeasy Blood and Tissue Kit from Qiagen (Hilden, Germany) was used for DNA extraction. Samples were further QC’s and sequenced at Novogene (Beijing, China) following the standard Illumina protocols. Library preparation produced fragment libraries of ¡350 bp insert length. Sequencing was done on Illumina HiSeqX-Ten to collect 150 bp paired-end data.

In this report we abbreviate the strain names by the first two letters: AC, BN, BU, F3, M5, MR, WK and WN.

5.2.2 Sequence alignment and variant calling

We aligned the raw sequence reads to the rat reference genome (rn6) using BWA version 0.5.9 [84], removed duplicates using Picard v1.76 [18], and performed realignment, recalibration and joint variant calling across eight strains with the UnifiedGenotyper with GATK v3.4. We removed variant sites with fewer than 10 reads in eight samples, and variant site quality score (QUAL) <= 30. We chose not to use the HaplotypeCaller as we have only eight inbred lines, which are not the population-based samples suitable for building haplotypes.

ChrX data showed the same pattern of heterozygosity as the autosomes in all eight animals, thus confirmed that they are female. We excluded the Y chromosome calls in downstream analysis. We did not call indels in this data release.

For comparison purposes we also ran the analysis with (1) two earlier versions of the reference genome, rn4 and rn5; (2) two other aligners. The first is by feeding the BWA aligned files into Stampy [90] version 1.0.32. Stampy alignment shows higher sensitivity than BWA, especially when reads include sequence variation. (The use of BWA-alignment as input for Stampy is to increases alignment speed without reducing sensitivity). The second is Bowtie2 v2.1.0 [80]. All the post-alignment processes followed the same Picard and GATK steps.

5.2.3 Comparison with the previously published variant calls using SOLiD

We compared our variant call set (for BWA alignment and rn6) with that by Hermsen *et al* [61], which was based on the SOLiD sequencing data. As that call set was aligned to rn5, we lifted over the variants to the rn6.

5.2.4 Defining regions of unusually high rates of heterozygosity

The final call set contains >16.4M SNV sites. We divided the genome into 1000-site windows, with a median window length of 221,100 bases. For each of the eight samples and in each window we computed the fraction of heterozygous sites in that window (only using the number of non-missing sites in that window as the denominator). Based on the inflection point in the empirical distribution of this per-window heterozygosity fraction (Figure 5.1) we chose a cutoff of 25% to designate a windows as of high-heterozygosity. We concatenated adjacent windows of high-heterozygosity into the same segment, and in a second step, merged adjacent high-het segments if they are separated by a single "low-heterozygosity" window, if that window had more than 0.175 heterozygote rate. After merging, there is no evidence of very short low-het segments separating high-het segments Figure 5.7

5.2.5 Data records Released

We share data records at different levels of processing.

1. Raw FASTQ/BAM files will be deposited in a public repository.
2. Variant calls from the UnifiedGenotyper (using BWA and rn6), as VCF files for the eight strains.
3. Mask files on our GitHub page13, documenting the regions of high heterozygosity, for each of the eight founder strains, as well as regions that are highly-heterozygous in all of the eight.

5.3 Results

5.3.1 Description of variant calls

Median read depth over the genome ranges 24X - 28X across the eight samples. Joint variant calling revealed 16,405,184 post-filter single-nucleotide variant sites on the autosomes and chromosome X. The number of heterozygous sites per line varies from 1,560,708 (BN) to 2,114,990 (WKY) (Table 5.1). BN represents the reference genome, and had more Ref/Ref than Alt/Alt genotypes. In contrast, the other seven strains had a comparable number of Ref/Ref calls as Alt/Alt calls.

We compared our genotype calls with those from Hermsen *et al* by calculating between-study concordance rates at sites reported in both, and using genotypes that do not include the missing calls. Table 5.2 shows that each of eight lines can be correctly matched between the two datasets, confirming the sample identity even when the two studies were based on different animals for a given line. BN had the highest between-study concordance: 0.95. Six other lines have concordance > 0.86. However, MR had the lowest concordance, 0.69. To determine if an animal from another line was mislabeled as MR, we compared our MR data with a larger panel of 42 strains previously published and found that our MR had the highest match with MR and WAG-Riji in that study. This suggests that the MR lines in different laboratories may have diverged to an unusually large degree.

Variant calls from the three aligners (BWA, Bowtie2 and Stampy performed comparably–we show genotype concordances in Table 5.3 (and concordance between high-heterozygosity segments in 5.4)).

5.3.2 Regions of unexpected high-heterozygosity

The eight founder strains are inbred over many generations, and thus we expect low heterozygosity in each strain across the genome [**?**]. However, when we calculated the fraction of heterozygous genotypes in consecutive 1000-SNV windows, we observed highly varied distribution of this metric along the genome. Not only there were many windows of high heterozygosity (>0.25), they also tended to recur in multiple lines (Figure 5.2). Some of these windows were found in all 8 strains, including BN, the strain of the reference genome 5.3.

We chose a cutoff value of 25% according to the distribution of per-window heterozygosity (Figure 5.2). After merging neighboring high-het windows if they are separated by a single low-het window with het rate > 0.175, we obtained 304-482 contiguous high-het regions across the eight lines (median 452), covering 176.4-254.8 Mb, equivalent to 6.3%-9.2% of the genome (average 8.4%). The distribution of segment lengths is shown in 5.4. Of the heterozygous calls in each line, 28-31% fall in the high-het regions (mean 29%).

The high-heterozygous regions found in all eight lines contain 1,756 Ensembl genes. These genes were enriched for G-protein coupled receptors and olfactory receptors (fold enrichment of 3.2 and 2.0 respectively), which are known to have many paralogous copies in mammalian genomes [67]. There are 4,963 missense variants, 123 stop-gain variants, and 154 splice donor/acceptor variants in these regions, belonging to 372 unique genes.

5.3.3 Heterozygous calls tend to be recurrent and show higher read depths

While Figure 5.1 shows that 1000-variant windows of high heterozygosity tend to appear in multiple lines, we also analyzed individual heterozygous genotypes to see how much they tend to recur in multiple lines. We divided the 16.7M variant sites according to the number of heterozygous genotypes observed in the eight lines, thus defining nine variant categories, from 0 to 8 heterozygotes (Figure 5.5). If the heterozygous genotypes appear independently with a probability of p, the expected chance of seeing a site with two heterozygotes (that is, in two of the eight lines) would be proportional to p^2, and in three lines: p^3. We estimated the upper limit of p by counting the fraction of heterogeneous genotypes over all 8 lines in all sites, knowing that this fraction is already biased upward due to the highly recurrent heterozygous sites. The expected probability of seeing k heterozygous sites, a(k)pk, where a(k) is the sampling coefficient of the binomial distribution, drops much faster than the observed k-het counts (Figure 5.5). For instance, the observed number of sites that are heterozygous in all eight lines is more than five orders of magnitude more than expectation.

The tendency for an individual site to appear heterozygous in multiple lines is related to the read depth at these sites, as we see in Figure 5.6. Notably, the read depth is higher in these regions for both homozygous and heterozygous genotypes.

5.3.4 Concordance with previous calls

We analyzed the previously released variant calls from Hermsen *et al* 2015 (after lifting over these calls to the rn6 genome from the original rn5) to see if our high-heterozygosity segments were also found in their calls or if it could be an artifact of our sequencing pipeline. We find similar regions showing up as heterozygous in both calls (Figure **??**), with an overlap of as much as 36% as defined by the length of the intersect of segments called high-heterozygosity in both calls, divided by the union of regions defined as high-heterozygosity in either call.

We have compiled a list of these low-confidence regions of the genome and provided them as bed files on our Github page: http://github.com/shwetaramdas/rataccessibleregions/.

5.4 Discussion

We present new evidence that the existing reference genome for rats has problematic regions which likely represent regions of mis-assembly. These regions make up ˜8% of the genome, and have a significantly higher read depth than other regions. A mixed approach using BACs, PacBio long-read sequences and Illumina short-read sequencing was used to assemble the rat reference genome. Since this was the first organism for which this approach was used, it was not entirely clear how this approach would perform in separating regions of high sequence similarity, such as segmental duplications.

Guryev et al1[55] in 2008 studied the rat reference genome available at the time (rn4) and used the read-depth distribution to identify their estimates of genome mis-assembly. These regions make up ˜1% of the genome, and only 2 of these 73 regions lift over to the rn6 genome (both these regions are called high-heterozygous in our data). However, they didn't take into account heterozygosity, which may resulted in their having under-called potentially mis-assembled regions. Another study on another rat inbred strain SHR/Olalpcv used short-read Illumina sequencing to obtain the genome sequence of this strain. This group also observed higher-than-expected levels of heterozygosity, and suggested that this could have resulted from collapsing of reads from segmental duplication17. Because they worked on a single strain, they could not observe the consistency of these mis-aligned regions across strains, and did not quantify/tabulate these regions.

Previous work on the mouse genome has shown that segmental duplications were inadvertently misassembled in a draft of the mouse genome [11]. Our estimates of misassembled regions thus represent the most comprehensive record of sequence problems with the rat reference genome. These regions harbor more than 2,000 genes, and studies failing to take notice of the poor quality of these regions are in danger of identifying incorrect candidate genes for their studies. We propose that the rat community mask out these regions while performing genomic analysis on rats, until a reference genome of higher quality is made available.

Figures

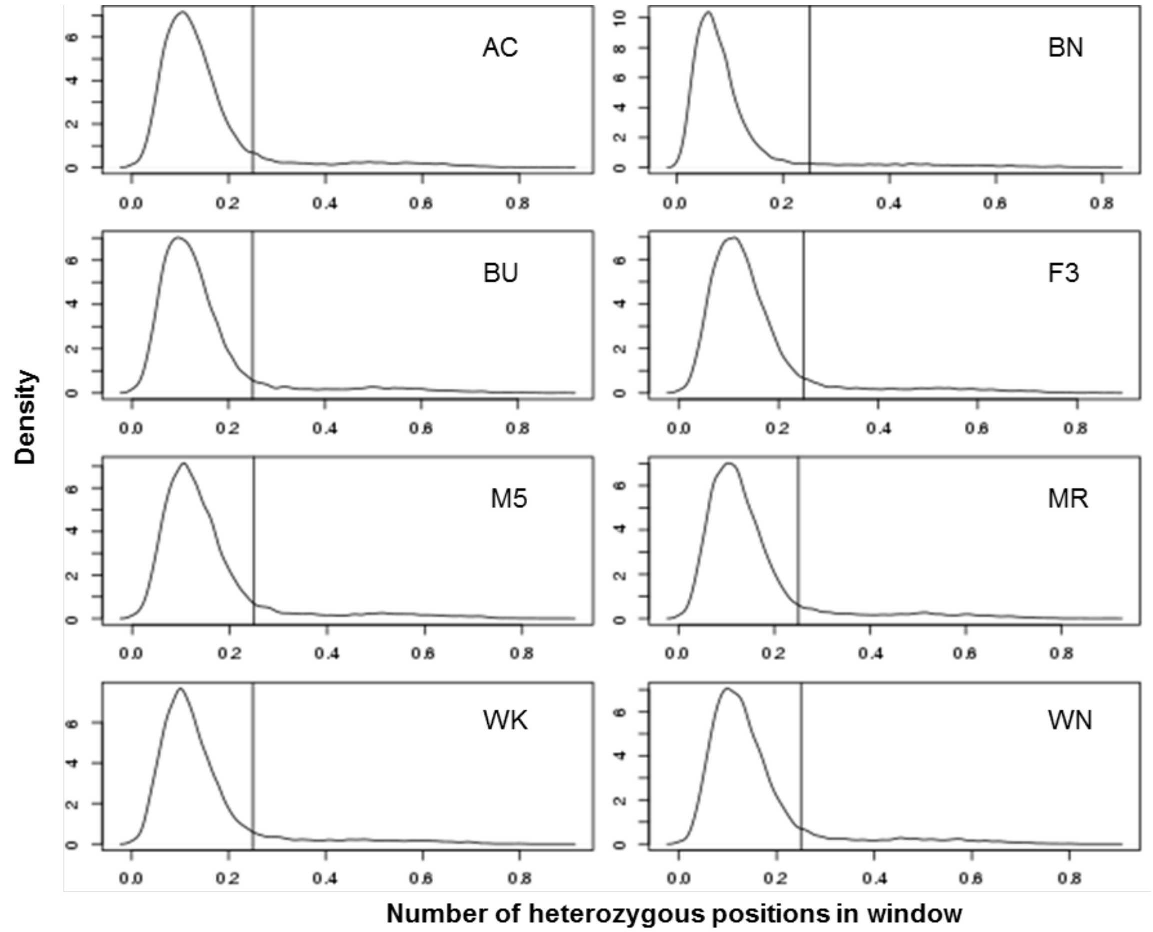

Figure 5.1: **Distribution of heterozygosity in 1000-marker windows across the genome** Distribution of the heterozygosity level in 1000-SNV windows for each of the eight lines, showing that 0.25 is a reasonable cutoff for defining high-het windows.

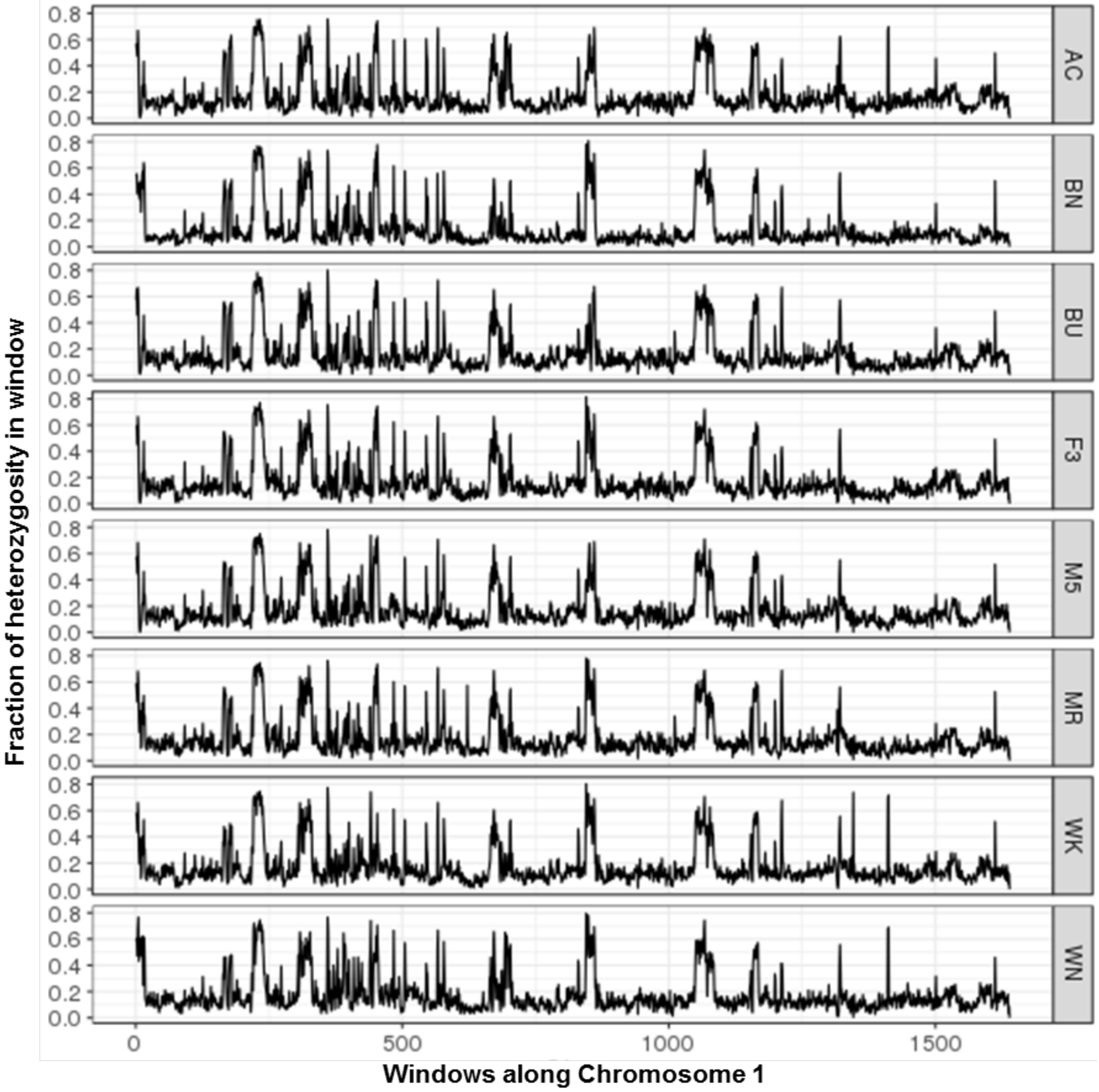

Figure 5.2: **Heterozygosity in Chromosome 1** Consistent heterozygosity patterns across the eight lines. Shown are the fraction of heterozygous genotypes in non-overlapping 1000-SNV windows in chromosome 1, displayed for the eight inbred lines.

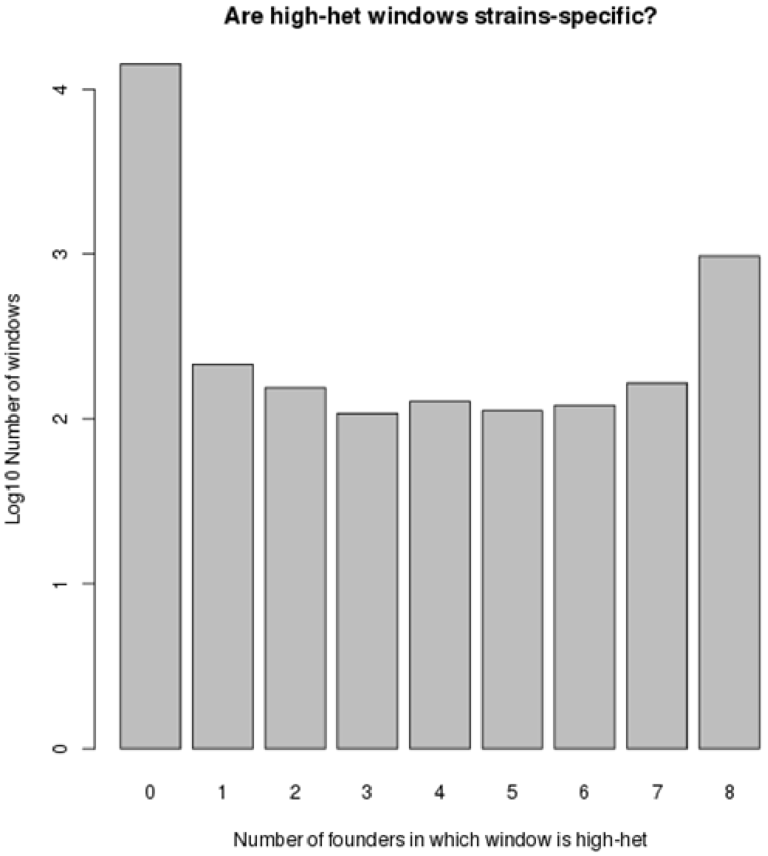

Figure 5.3: **High heterozygosity windows have more hets** Windows of high heterozygosity (bottom panel) correspond to markers with higher read depth across Chromosome 1 for strain AC (similar results are seen for other chromosomes and other strains)

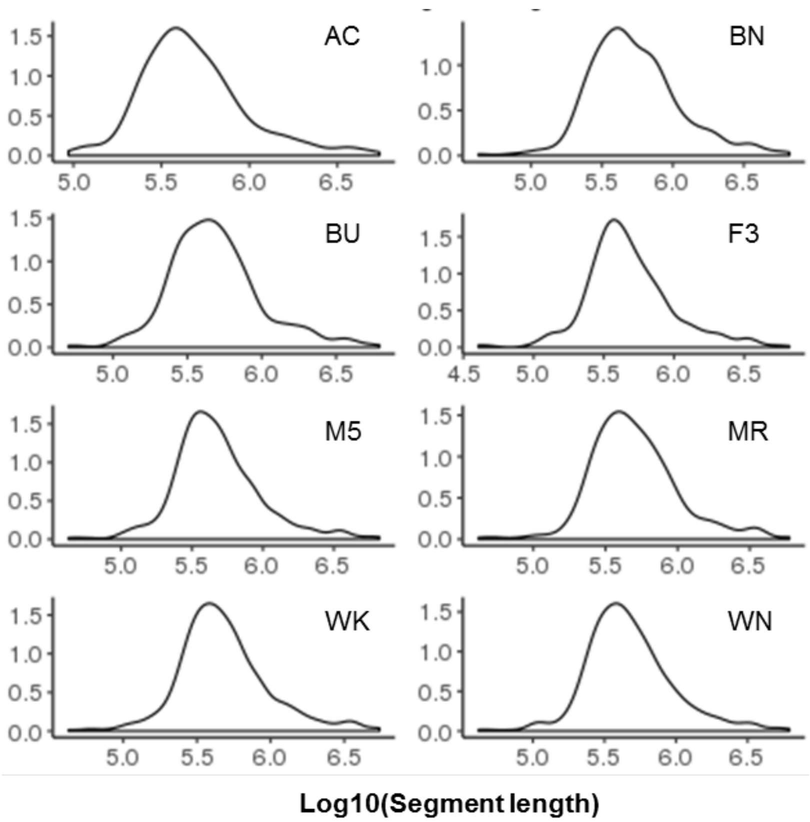

Figure 5.4: **Distribution of segment lengths in eight founders** Shown is the distribution of lengths of high-heterozygosity segments in each of the eight founder strains. We see that each of the eight strains has a similar range of segment lengths, with a mean of ˜700Kb.

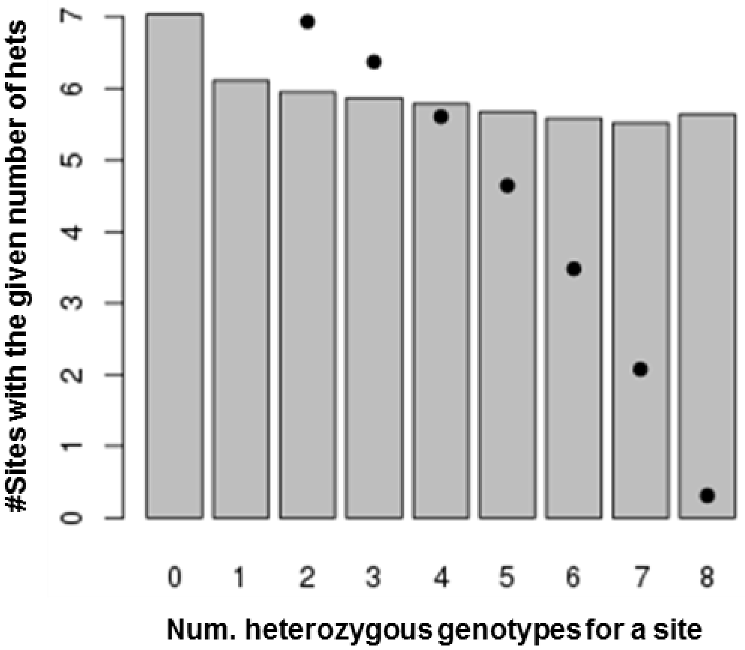

Figure 5.5: **Heterozygosity distribution per site across 8 founders.** Recurrence pattern of heterozygous genotypes across 8 founders. All variant sites were categorized as having 0, 1, , or up to 8 heterozygous genotypes. The bar graph displays the histogram of the observed sites, while the dots show the expected number of sites if recurrence is random, estimated under a simple binomial model. The expected values for 0 and 1 founders exceed the limits of the plot.

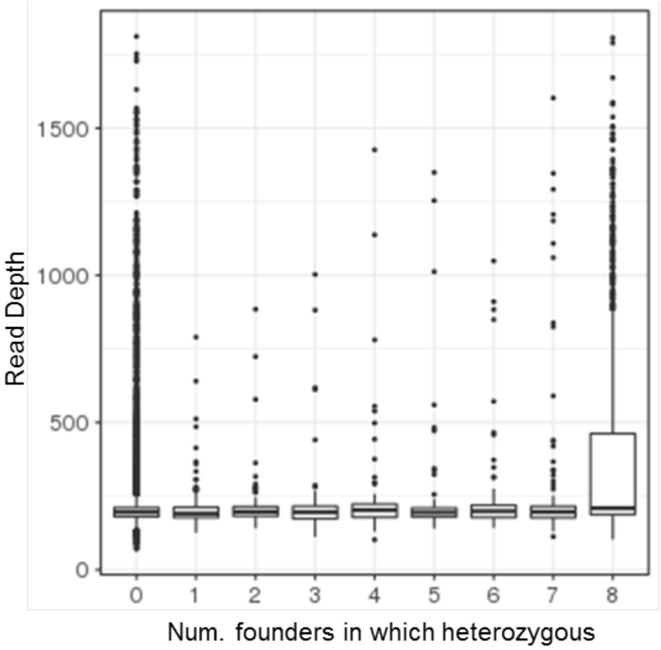

Figure 5.6: **Heterozygous windows show higher read depth** Windows heterozygous across all eight founders (right-most bar) show a significantly higher read depth than windows heterozygous in 0-7 founders

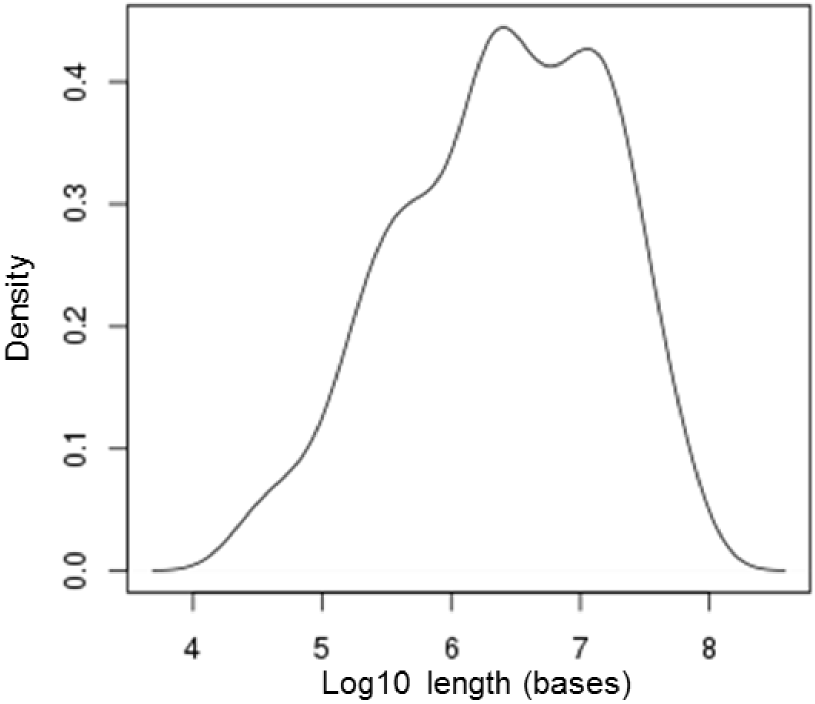

Figure 5.7: **Length distribution of low-heterozygosity segments** After merging segments separated by a single low-het window with heterozygosity > 0.175, we see there are not extremely short low-heterozygosity segments.

Tables

Strain	**0/0**	**0/1**	**1/1**	**./.**	**Others (>2 alleles)**	**Het frequencies**
AC	6,249,928	2,096,993	5,421,143	2,385,991	251,129	0.150
BN	12,324,354	1,560,708	803,800	1,623,744	92,578	0.106
BU	6,688,095	2,058,214	5,192,072	2,218,374	248,429	0.145
F3	6,587,382	2,101,058	5,282,082	2,181,114	253,548	0.148
M5	6,628,579	2,160,032	5,329,841	2,031,084	255,648	0.150
MR	6,736,380	2,111,216	5,243,452	2,063,655	250,481	0.147
WKY	6,599,091	1,976,801	5,130,247	2,456,145	242,900	0.142
WN	5,993,693	2,114,990	5,734,351	2,304,637	257,513	0.150

Table 5.1: **Genotype frequencies in each of eight founder strains**

	Hermsen 2015 calls								
		AC	BN	BU	F3	M5	MR	WK	WN
Our variant calls	AC	0.874	0.478	0.577	0.588	0.584	0.599	0.514	0.547
	BN	0.52	0.952	0.548	0.541	0.55	0.559	0.497	0.551
	BU	0.586	0.511	0.871	0.628	0.631	0.609	0.52	0.599
	F3	0.594	0.501	0.625	0.873	0.674	0.593	0.52	0.594
	M5	0.591	0.511	0.63	0.676	0.873	0.601	0.514	0.592
	MR	0.585	0.521	0.605	0.599	0.598	0.685	0.526	0.603
	WK	0.518	0.453	0.516	0.518	0.51	0.536	0.859	0.508
	WN	0.55	0.511	0.594	0.591	0.588	0.599	0.508	0.88

Table 5.2: **Concordance of genotypes between our variant calls and those from Hermsen et al 2015**

	BWA	**Bowtie**	**Stampy**
BWA	1	0.972	0.972
Bowtie		1	0.974
Stampy			1

Table 5.3: **Genotype concordance between aligners**

	BWA	**Bowtie**	**Stampy**	**BWA: Bowtie**	**BWA: Stampy**	**Bowtie: Stampy**
AC	238,523,232	260,711,962	302,820,837	0.57	0.54	0.51
BN	171,689,865	189,496,386	213,984,578	0.61	0.6	0.58
BU	223,339,767	247,060,856	289,737,173	0.55	0.55	0.49
F3	236,379,041	252,011,886	293,848,725	0.56	0.54	0.51
M5	243,992,401	243,643,168	294,212,048	0.58	0.54	0.5
MR	225,952,435	263,629,202	276,523,184	0.55	0.55	0.5
WK	242,849,594	239,016,453	288,897,012	0.57	0.54	0.52
WN	228,435,487	249,395,735	298,824,484	0.53	0.52	0.47

Table 5.4: **The length of heterozygous segments across the genome, compared across 3 aligners. Bowtie gives the least heterozygosity, followed by BWA, and Stampy shows the worst heterozygosity**

CHAPTER VI

A Framework for Power Calculations in CNV detection from Sequencing Data

6.1 Background

A common approach for CNA-detection relies on the telltale stepwise change of apparent DNA ”dosage” in a genomic region [106]. For data from microarray-based comparative genomic hybridization (arrayCGH) or single nucleotide polymorphism (SNP) genotyping, the raw signal for DNA dosage is the total hybridization intensity at the individual locus, which is either the arrayCGH probe location or the SNP site. The distribution of these measured locations along the genome has been determined by how the genetic markers were selected during the design stage of each array platform. A more recent technology is high-throughput DNA sequencing, where the observed number of mapped reads in a given genomic interval reflects the regional variation of DNA in the input material. In this study we mainly consider sequencing data, and do not make the distinction between germline DNA or somatic DNA, and we will use CNA and CNV (copy number variation) interchangeably. When discussing sample purity, we refer to either the mixing of two germline DNA samples or the coexistence of two somatic cell populations (such as cancer cell clones). We assume that all the mapped reads are correctly mapped, thus sidestepping the situation when segmental duplications might cause some reads to be ambiguously mapped to multiple locations. We will not address the situation of copy-neutral structural variation such as loss-of-heterozygosity or inversions.

For both microarray data and sequencing data, regional variation of signal intensity along the genome presents a major challenge for detecting step-wise changes. Systematic biases, such as those due to local GC-content, are technology-dependent and can be largely corrected by appropriate normalization based on empirical bias patterns learned from a large sample [34]. In this study we will not dwell on the issue of bias correction, and will assume that the best procedures for correcting systematic variation will have been applied in early stages of data preparation. Instead we will return to the topic of signal over-dispersion, i.e., the ”noise” of read depth along the genome beyond what is expected for random sampling from a Poisson distribution.

Read depth-based CNV detection can be formulated as a task of comparing of the mean between two samples, as implemented in parametric tests such as the Student’s t-test. Importantly, the unit-of-observation depends on the technology. For microarrays, the ”data point” is the quantitative intensity of an arrayCGH probe or the total signal of the two alleles at a SNP site. Here the units, and their spatial distribution, have been pre-defined by the chosen platform; and in general, the sampled locations tend to cover the genome at variable intervals. For whole-genome sequencing (WGS) data, the analyst can choose a window width and tally the total number of reads in consecutive windows, where each window provides a data point. If a CNA contains n data points, the task is to determine if the mean of such n observations is different from the mean of a null distribution, which can be established either with adjacent ”normal” intervals or with a genome-wide baseline.

We denote the haploid sequencing read depth as D (for a dataset of 30X total coverage, the haploid coverage is 15), the length of the CNV as L, the size of the window into which markers/reads are binned as W, the ploidy of the CNV as N (normal diploid regions have N=2), read length is l, and purity of the sample as F, defined in this study as a two-way mixing of a normal sample at (1-F) and an aneuploidy sample at F. We do not consider three-way or more complex mixing.

If we ignore the reads falling on window boundaries, a window of length W would contain W/l reads laid end-to-end. At a depth of D this window is expected to contain X=2DW/l reads in a diploid region, and NDW/l reads in a region of ploidy N. When we use the mean (Avg(X)) to estimate the expected per-window read count, $\bar{X}$, a t-like statistics can be constructed as the difference of the mean between the N-ploid CNA and a diploid baseline, scaled by the standard error of this difference:

$$t = \frac{Avg(X1) - Avg(X2)}{\sqrt{var(X1)/n1 + var(X2)/n2}} \tag{6.1}$$

where X1 and X2 are the per-window read count in the CNA and outside the CNA, respectively. n1 and n2 are the number of independent measurements, i.e., the number of windows, in and outside the CNA, respectively. The numerator can be expressed as $|N-2|$DW/l. A CNA of length L contains n1=L/W windows. Next, we make the assumption that most of the genome is diploid and the baseline diploid regions can be pooled such that n2 is much larger than n1. By omitting the (1/n2) term the t-like score is simplified to

$$t = \frac{|N-2|DW/l}{sd(X_1)}.\sqrt{L/W} \tag{6.2}$$

where $X_1 = NDW/l$

where the denominator is the standard deviation of X=NDW/l in the CNA region. The last term is $\sqrt{n1}$, the square root of "sample size", i.e., the number of windows in the N-ploid CNA.

We would like to re-emphasize that we used D to denote haploid read depth in order to make the expression less cluttered. For instance, in an experiment with mean coverage of the diploid genome of 30X, a region with a ploidy of 3 would be expected to have a coverage of 45X (with an expected difference of 15).

The sample size n here (the number of independent events supporting the CNV) can be written as n = (L/W), the number of bins in the CNV.

The t-like score in Eq-1 and Eq-2 represents the total signal in a power analysis, to be partitioned into two components. $t = t_\alpha + t_\beta$ where t_α is the component for a significance level of α, and t_β is the component for a power of 1-β. In essence, the total t-like score is to be "spent" in two ways, one for a specific Type-1 error rate, and the rest for a specific Type-2 error rate. In the rest of the manuscript we will derive the expressions for the total t score, whereas in depicting the power curves we fix t_α for a per-event Type-1 error of 0.05. In real situations, the experiment-wide Type-1 error needs to be calculated with multiple testing correction according to the number of CNAs to be detected for the entire sample.

6.2 Motivation

As outlined above, a standard power analysis involves five key components: on one side of the expression is the total weight of evidence, split into t_α+t_β, to account for the tradeoff between the false positive and false negative rates, On the other side of the expression is the effect size, divided by the standard deviation of this effect, and multiplied by the square root of the sample size. In the context of coverage-based CNA detection, the effect size is a function of four parameters: |N-2|DW/l, sample size is L/W, proportional to L.

This study is motivated by the need to explore three factors that affect power. First, the choice of W brings two opposing consequences: a larger W increases effect size but reduces sample size. We will examine the net outcome of varying W. Second, the number of reads in consecutive windows may not vary according to a simple Poisson sampling process. Even after correcting for systematic regional biases, certain levels of over-dispersion may remain, leading to higher values in the denominator and reduced power. We will introduce a variance inflation factor in the power calculator. Three, many samples contain the mixing of two populations, and sometimes the CNA-bearing population is very rare. Can a 10-fold reduction of ”purity” be compensated by 10-fold deeper sequencing. We will examine questions like this, with or without the variance inflation factor. The results below are organized into three main parts. In #4 we examine the situation of 100% purity (F=1), with or without variance inflation. In #5 we introduce F¡1, again with or without inflation. In #6 we explore two use cases, one with very long CNAs (large L) but low sequencing depth (small D); the other with high sequencing depth (large D) but low purity (small F).

6.3 100 percent tumor purity

6.3.1 Poisson

In the simplest case, read counts from sequence data can be modeled as a simple Poisson [reference] where the mean equals the variance. The distribution of X_1, the read depth in the region of the CNV, a Poisson with mean (N)*D*(W/l). Then, the t-statistic from Equation (1), using mean values from Equation (2) becomes

$$t_1 = \frac{|N-2|}{\sqrt{N}} \frac{DW/l}{\sqrt{DW/l}} . \sqrt{L/W} \tag{6.3a}$$

$$t_1 = \frac{|N-2|}{\sqrt{N}} \sqrt{\frac{DL}{l}} \tag{6.3b}$$

From this equation, we see that the power of the test is directly proportional to the read depth (D), and the length of the underlying CNV event (L), and inversely proportional to the length of the read (l). All of these relations are consistent with expectations, but expressed quantitatively.

In particular, we notice that the t-statistic in this scenario is not influenced by our choice of window size W. While a large W reduces variance, it also reduces the effective sample size, and these two opposing effects balance out each other in the calculation of the t-statisic. However, the *power* of the test is still dependent on the window size through the number of degrees of freedom. As n $>$ 5 , i.e. with large CNVs, using multiple windows to tile them, this dependence of power on n disappears. In Figure 6.1, we fix N(3), l(100), W(1000), and show power as a functions of length of the CNV (L) and the read depth (D) under a Poisson model. We see that for a CNV that is 10Kb long, read depth of 0.1X shows very low power while increasing read depth to 1X increases power considerably.

6.3.2 Negative binomial

Sequencing experiments were initially expected to follow a Poisson distribution, but studies have shown that HTS experiments show an overdispersion in variance, which is often modeled using a negative binomial distribution (Love, 2013). Wang *et al* 2008 [139] showed that the variance in their whole-genome data from a single individual was twice the read-depth. Using this estimate of dispersion, we use $\phi = 1/\mu$ as the default over-dispersion parameter.

This over-dispersion depends on the sequencing technology used due to biased amplication or sequencing of different regions of the genome. In general, exome-sequencing has been shown to have greater variability than whole-genome approaches due to additional bias from the targeting approaches.

We can model our X_1 as a negative binomial distribution with same mean as the Poisson but an added term to the variance.

$$Variance(\Delta) = \mu(1+\mu\phi)$$

where μ is the mean of the Poisson (as before, μ in a region of ploidy N is N*W*D/l), and ϕ is an overdispersion parameter estimated from the sequencing data.

The t statistic from Equation (6.3b) now has an additional parameter from the over-dispersion:

$$t = \frac{|N-2|}{\sqrt{N}}\frac{DW/l}{\sqrt{DW/l}}.\sqrt{L/W}.\frac{1}{\sqrt{1+\mu\phi}}$$

As before, the N/2 term represents the fold-increase in reads at a region with ploidy N compared to a diploid. For instance, in an experiment with median coverage 30X, a region with a ploidy of 3 would be expected to have a coverage of 45X.

The equation above simplifies to:

$$t = \frac{|N-2|}{\sqrt{N}}\sqrt{\frac{DL}{l}}\frac{1}{\sqrt{1+\mu\phi}} \tag{6.4}$$

this equation doesn't simplify without assumptions about the parameter ϕ.

The t-statistic above is the t-statistic from the Poisson model with an extra variance parameter.

Keeping constant the read length l, the window size W and the ploidy N, we estimated the effect of the over-dispersion $\mu\phi$, read depth D and CNV length L (Figure 6.2). We find that within reasonable estimates of $\mu\phi$, the variance increases with window sizes similar to a Poisson distribution, and in cases where users expect large CNVs, a Poisson model can be fit to the data without much loss of information.

6.4 Taking into account tumor impurity

Tumor purity is the proportion of cancer cells in the admixture of cells in the tumor microenvironment. This includes cancerous cells and non-cancerous cells, like fibroblasts, epithelial cells, and immune cells [7]. In recent years, sequencing of blood or plasma DNA (cell-free DNA or cfDNA) has been used to detect cancer-derived mutations and CNVs. It has been demonstrated that it is feasible to detect somatic CNVs using sparse whole-genome sequencing with 0.1X coverage [60].

When we take this tumor impurity into account, our power calculations are modified by the additional parameter F (tumor fraction).

The t statistic is:

$$t = \frac{F|N-2|DW/l}{\sqrt{Var(FX_1 + (1-F)X_2)}}.\sqrt{L/W} \tag{6.5}$$

where, as before, X_1 is the read depth in the region of the CNV (N*DW/l) and X_2 is the background read depth in the genome (2*DW/l).

With the assumption that F $\ll$ 1,

$$t = \frac{F|N-2|DW/l}{\sqrt{Var(X_2)}}.\sqrt{L/W} \tag{6.6}$$

6.4.1 Poisson

Under the assumptions of the Poisson model, without assuming low sample purity, this becomes

$$t = \frac{F|N-2|DW/l}{\sqrt{Var(FX_1 + (1-F)X_2)}}.\sqrt{L/W} \tag{6.7}$$

which doesn't simplify.

To obtain a numerical estimate of the t-statistic from the above formula, we simulate values for the denominator, which can be represented as a linear combination of Poissons:

$$\text{Variance term in Denominator} = Var(FX_1 + (1-F)X_2)$$

Where X_1 is the Poisson distribution in the region of the CNV, and X_2 is the Poisson distribution in the background region. We can simulate this data to obtain a value for the variance.

Assuming low sample purity (Equation 6.6), we get:

$$t = \frac{FWD\sqrt{L/W}|N-2|/l}{\sqrt{WD/l}}$$

which simplifies to:

$$t = \frac{F\sqrt{LD}|N-2|}{\sqrt{l}}$$

We see that the t statistic is independent of the window size W, as in the case with 100% tumor purity. We plot the dependence of power on window size and read depth in Figure 6.3: we see for tumor fractions around 0.01, the power to detect even long CNVs is very low.

6.4.2 Negative binomial

WD/l follows a NB with over-dispersion parameter ϕ

$$t = \frac{F|N-2|DW/l}{\sqrt{Var(FX_1 + (1-F)X_2)}}.\sqrt{L/W}$$

We use simulations from the negative binomial distribution to estimate values for the denominator.

Under assumptions of low sample purity, the equation (6.6) becomes:

$$t = \frac{F|N-2|(DW/l)}{\sqrt{(2DW/l)1+\mu\phi}}.\sqrt{L/W}$$

which simplifies to

$$t = \frac{F|N-2|\sqrt{DL/l}}{\sqrt{2(1+\mu\phi)}} \tag{6.8}$$

We show, in **Figure 4**, power calculations for a range of read depths, keeping constant the tumor fraction F. Calculations for a range of parameters can be performed using our online calculator.

We now asked the question: how is power influenced by other experimental parameters, keeping tumor fraction constant at 0.1? From Figure 6.4, we first notice that power increases with increasing read depth D, and tumor fractions below 0.1 yield very low power for short CNVs (L less than 100,000, even at low values of over-dispersion).

6.5 Use Cases

We present some power calculations for some common experimental designs:

6.5.1 Shallow whole-genome sequencing to detect large CNVs in tumors (high tumor fraction)

Previous studies have shown that it's possible to detect large CNVs by shallow whole-genome sequencing (0.1X) of plasma DNA. We ran a power analysis using these parameters (N of 3, l 100, W 1000) in Figure 6.5, and show that the power to detect CNVs smaller than 10Mb is very low with a read depth of 0.1X (and for tumor fractions less than 0.1, we need very high read depths to reach this power).

6.5.2 Targeted deep sequencing of previously known CNV in samples with low tumor fraction

To check for recurrence of a tumor, patients can choose to sequence extracellular DNA at regular intervals. We can use a power analysis to optimize frequency of sampling, given a fixed cost. Figure 6.6 shows power calculations for this scenario, to identify relatively short CNVs of length 10Kb (other fixed parameters include l 100, N 3 and W 1000). We see that low read depths fail to yield high power, even at higher values of tumor fractions F. We also see in the Poisson model (data not shown) that the increase in power from a 10-fold increase in the length of the CNV L is equivalent to the decrease in power from a 10-fold reduction in read depth.

6.6 Caveats to this methodology

There are caveats to using this approach to power calculations. The first is that there are alternative approaches to CNV detection that take into account split reads [71], paired-end information [76, 47], single nucleotide polymorphisms, RNA-Seq data, or using a combination of these metrics [64, 81, 57]. Adding information from these data points would increase the power of CNV detection.

On the other hand, the sequencing read depth across the genome is affected by not just copy number, but many confounding variables. One such source of bias is the GC content of a region (which is known to have higher likelihood of being sequenced), another is the regional variation in the fraction of short reads that can be aligned to a given position. Whole-exome sequencing is subject to greater variation and confounding, including from capture efficiency of probes. Corrections for these confounders require sequencing information from multiple samples, and a failure to do so can lead to false positive CNV calls.

Figures

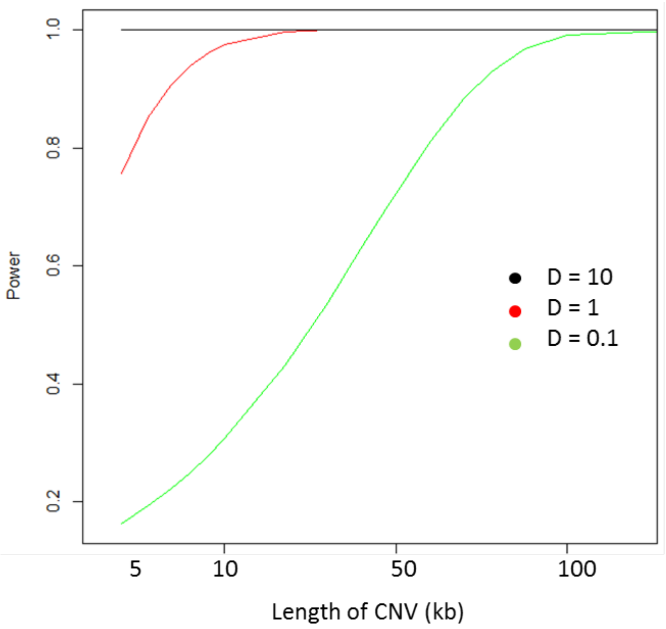

Figure 6.1: **Power analysis using a Poisson model with 100% purity** We plot the increase in power with increase in length of the underlying CNV L (x axis) and D (different colors), keeping constant l (100bp), W (1000), and N (3)

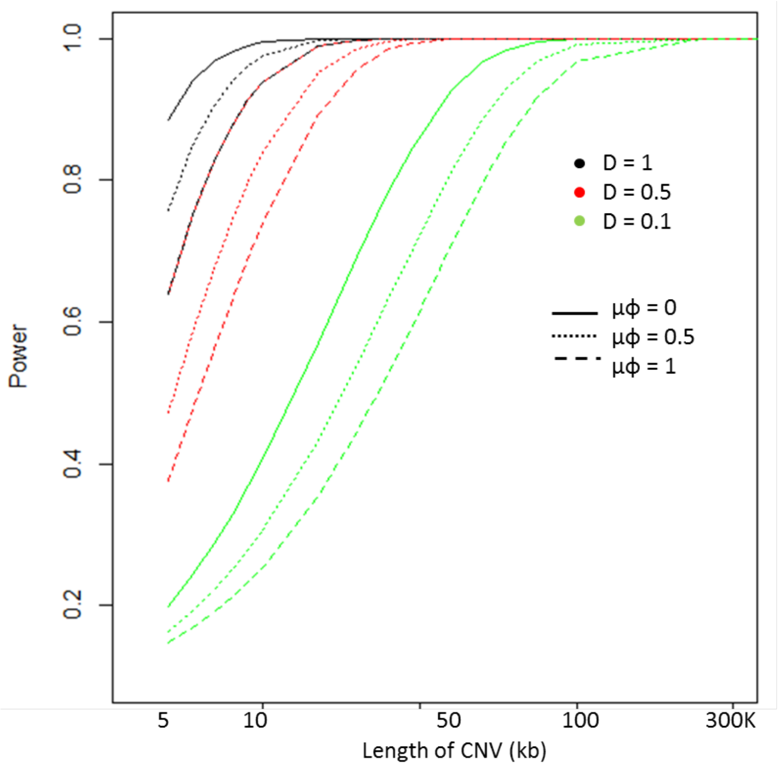

Figure 6.2: **Power analysis using a Negative Binomial model with 100% purity** We plot the increase in power with increase in length of the underlying CNV L (x axis), over-dispersion$\mu\phi$ (different colors) and D (different line types), keeping constant l (100bp), W (1000), and N (3)

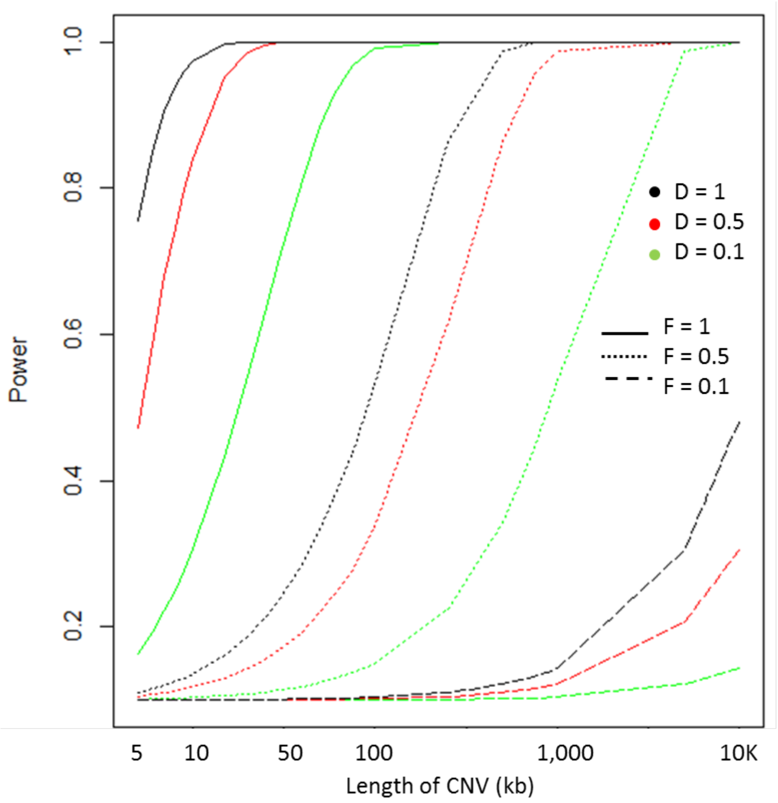

Figure 6.3: **Power analysis using a Poisson model with < 100% purity** We plot the increase in power with increase in length of the underlying CNV L (x axis), F (different colors), D (different line types) and over-dispersion (three panels), keeping constant l (100bp), W (1000), and N (3)

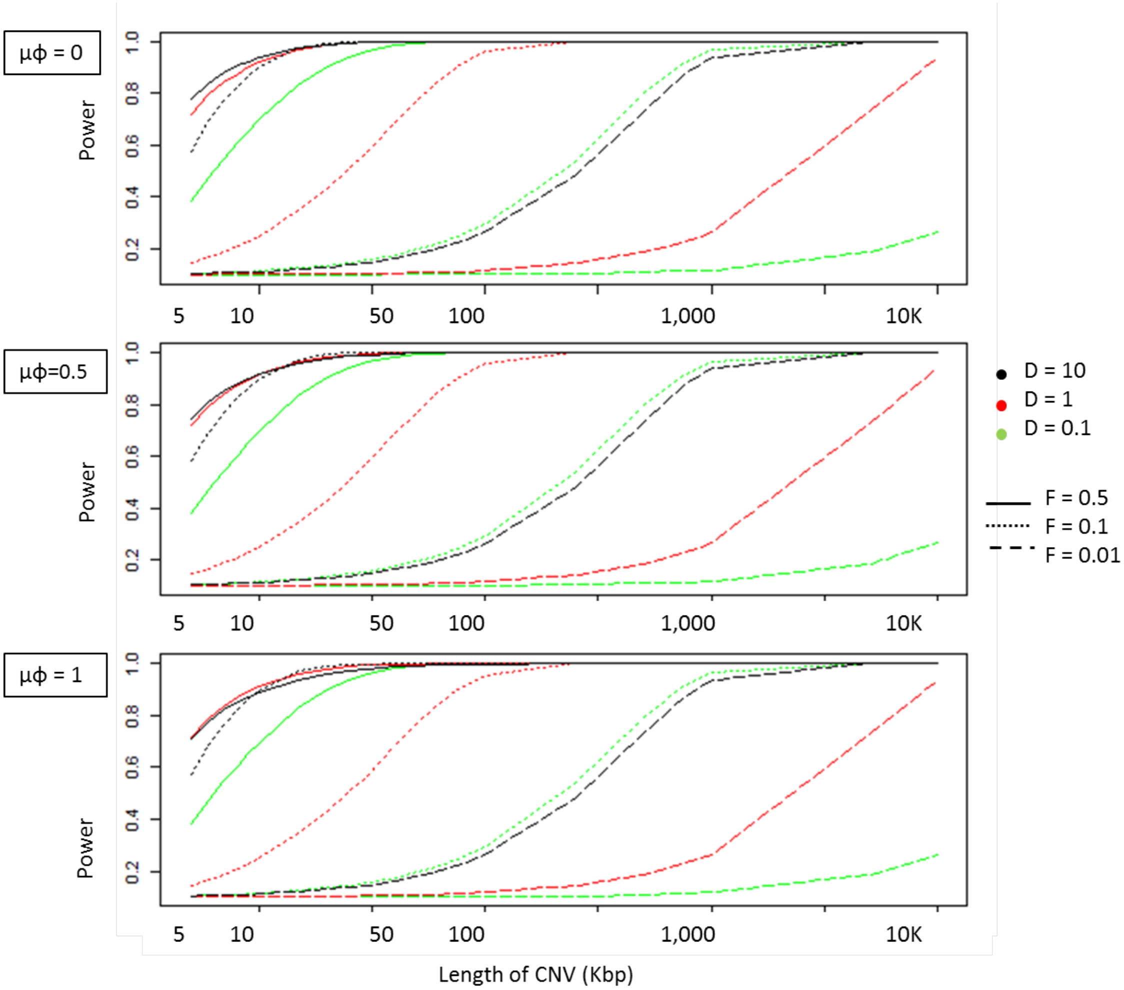

Figure 6.4: **Power analysis using a Negative Binomial model with < 100% purity** We plot the increase in power with increase in length of the underlying CNV L (x axis), F (different colors) and D (different line types), keeping constant l (100bp), W (1000), and N (3)

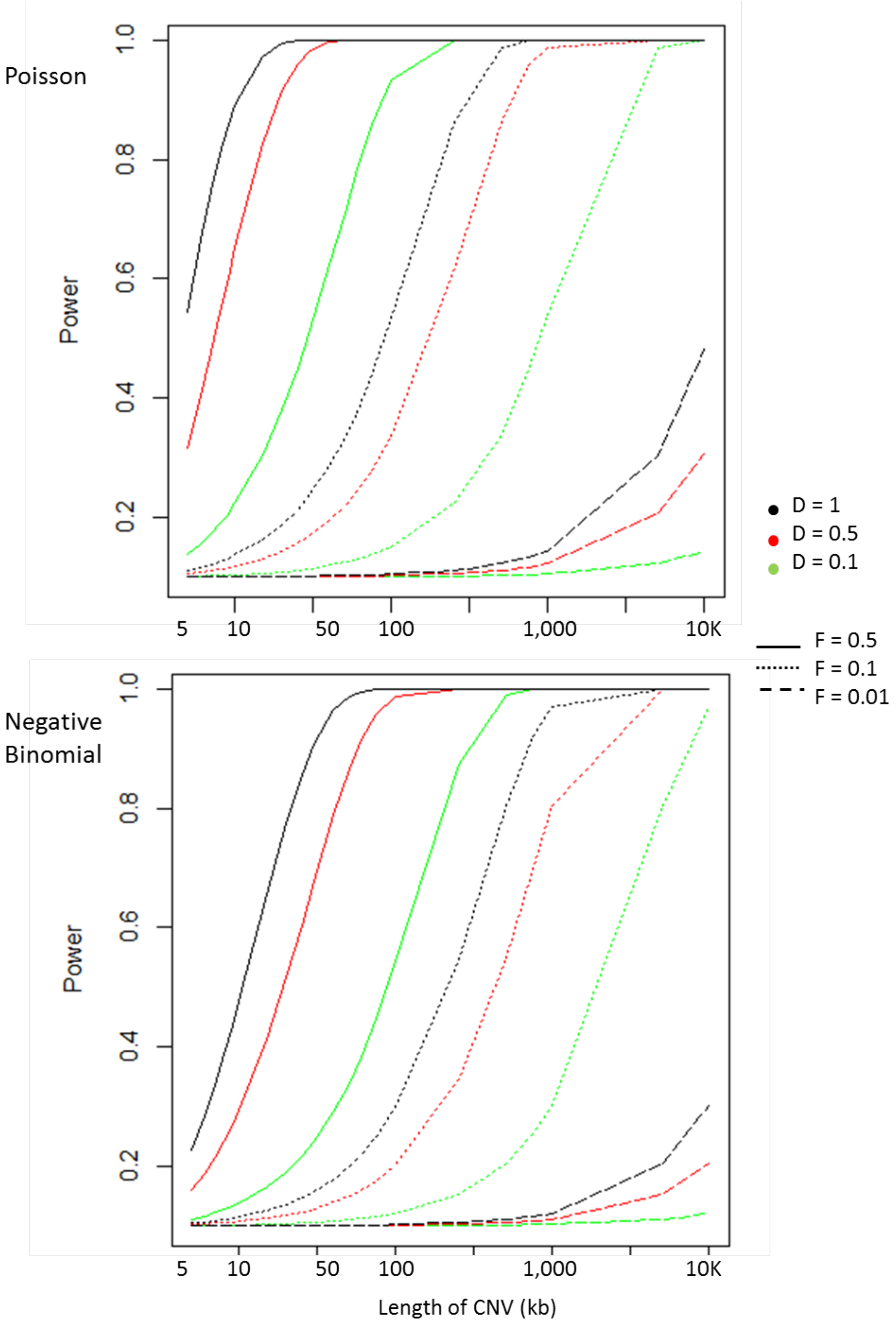

Figure 6.5: **Power Analysis for Use Case 1** Poisson model on the top panel, Negative Binomial model on the bottom panel

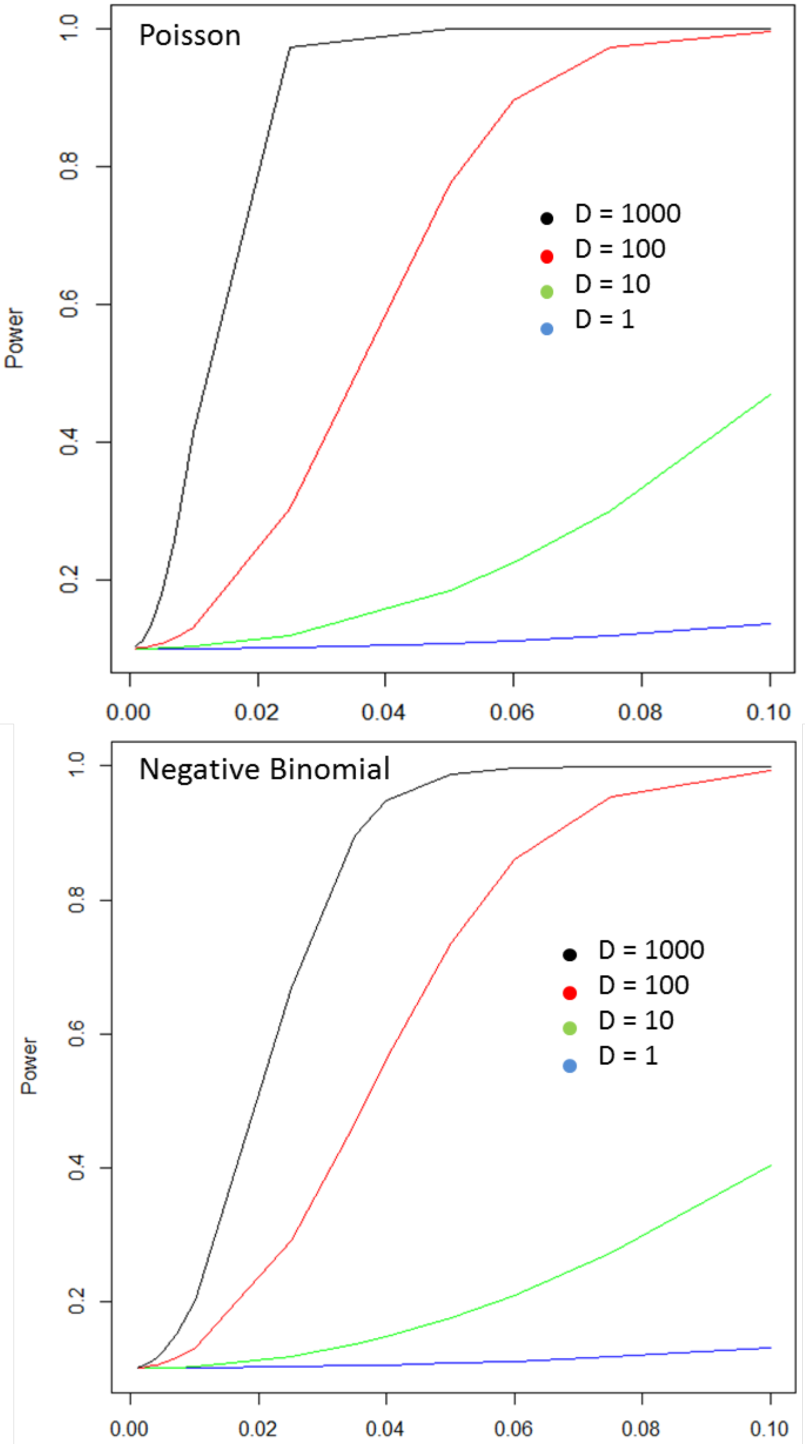

Figure 6.6: **Power Analysis for Use Case 2** Poisson model on the top panel, Negative Binomial model on the bottom panel

CHAPTER VII

Conclusions

If research in the field of complex genetics can be classified into biological discovery, tool development, and map-making/resource-building, this dissertation has made forays into each of these areas. In this dissertation, I have applied various complementary methods to understand the underlying functional basis of various traits. I have used genetic variation (Chapters 2, 4), as well as gene expression (Chapters 3, 4) to ask these questions. In Chapter 2, I used exome sequencing data from pedigrees with high incidences of bipolar disorder to attempt to identify causal mutations for the disease in these pedigrees. Our results indicate that bipolar disorder is likely to have a multi-factorial etiology even in these pedigrees with high incidence, and that there is limited convergence in genes across pedigrees. Our results do, however, point to a significant overlap with autism genes, and the implication of circadian genes. Another phenotype I studied is tissue aging, where, again, we see a polygenic signal, with many genes showing increase or decrease of expression with age.

In Chapter 3, I attempted to identify tissue signatures of aging in the trabecular meshwork using microarray data. We identified genes whose expressions increase or decrease with biological age, and several pathways implicated in the aging phenotype. We also saw that gene expression in the trabecular meshwork is a poor predictor of chronological age of the tissue (with performance failing to reach that of previous methylation-based predictors).

In Chapter 4, I integrated genetic data with gene expression data to map QTLs in a rat model of aerobic capacity. Even in a model with high selection pressure, reduced environmental variability, and a genetic background thats less diverse than that in human populations, our results are convergent with a genetic architecture thats polygenic. Our results highlight the role of fatty acid metabolic pathways and mitochondrial pathways, and point us to a small number of candidate genes supported by multiple lines of evidence for functional follow-up.

In Chapters 5 and 6, I worked on methods and resources of the complex trait genetics field. First, I analyzed the rat reference genome and identify mis-assembled regions in the assembly making up 8% of the genome, alerting researchers to the necessity of masking certain regions in the genome while performing genomic analyses; this is the first step in a concerted effort by the community to create a better reference genome for this model organism. In Chapter 6, I turned my attention to the detection of copy number variation from sequencing data—another class of variation contributing to variation in complex traits. Despite the many tools for calling copy number changes, there is still confusion about the proper investment of resources based on a principled power analysis. I created a new framework for power calculations to detect copy number variations from sequencing data, taking into account various experimental parameters affecting these. I also created an online calculator for researchers and clinicians to estimate the power of their study/biopsy under different experimental settings. As the use of sequencing technologies in the realm of clinical genetics continues to increase, an extension of this framework can maximize the utility of genetics in such settings.

The power of discovery of a genetic study is limited either by the technology and sample size (which determine what features we choose to shine a light on) or the strength of the underlying biological signal itself. Genome-wide studies are often called ‘hypothesis-free’, but this is not strictly true—while there is no biological candidate (like a gene of interest), the features these studies choose to look at determine the questions that are being asked, and set boundaries for what can be detected. Looking at exome sequencing data alone in Chapter 2 limited our search space to coding variants. If any of these pedigrees show regulatory defects leading to the phenotype, we will have missed these signals under our narrow hypothesis. The study on the trabecular meshwork missed splicing differences with age, because of the technical design of microarrays. Moreover, it is also possible that expression signatures in this tissue are particularly depleted of information (as seen by the weaker signal in this study compared to other similarly-powered studies on other tissues) because of low cellular density (and thus, fewer ‘live’ cells being assayed)—protein expression studies in this tissue may be a better indicator of age.

The results from my hunt for causal genetic signals for various traits (Chapters 2,3 & 4) point to a more diffuse non-Mendelian signal for the underlying trait than we had gone in expecting *a priori*, demonstrated by a lack of convergence across pedigrees in Chapter 2, and a deflated QQ-plot in Chapter 4. This has been the theme underlying research in the human complex traits field over the last decade. As a community, we have to respond to this new paradigm not by merely increasing sample sizes, but by focusing on strategies that take this complexity into account. Integrating information from regulatory data (reducing the number of tested interactions) is one way to better study these.

The contribution of the genetic background to complexity in mechanisms means that we need to think about the implications of results from model organisms in human complex trait genetics. A majority of model organism research (especially in mice and rats) has focused on minimizing both genetic and environmental variability—a principle that the use of inbred lines takes to an extreme. However, this means that while we can learn about biological mechanism from genetic studies using such models, the generalization of results from these models to identify pathogenic variants may be limited. This field may have to re-invent itself to become a better model for complex phenotypes in complex human populations.

When the human genome was first mapped in 2001, Bill Clinton described it as "the most wondrous map ever produced by humankind…today we are learning the language in which God created life". While we are well on our way to having read the script (and are beginning to find ways to edit it [25]), we are still understanding its grammar. In the next decade, we are likely to have multiple layers of information on the genome and the phenome. Integrating these judiciously will be a key tool in deciphering the algorithms of the genome.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101, 2002.

[2] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–249, 2010.

[3] L. Alonso, E. C. Souza, M. V. Oliveira, L. F. Do Nascimento, and P. M. Dantas. Heritability of aerobic power of individuals in northeast Brazil. *Biology of Sport*, 31(4):267–270, 2014.

[4] J. Alvarado, C. Murphy, J. Polansky, and R. Juster. Age-related changes in trabecular meshwork cellularity. *Investigative Ophthalmology & Visual Science*, 21(5):714–727, 1981.

[5] S. A. Ament, S. Szelinger, G. Glusman, J. Ashworth, L. Hou, N. Akula, T. Shekhtman, J. A. Badner, M. E. Brunkow, D. E. Mauldin, A. B. Stittrich, K. Rouleau, S. D. Detera-Wadleigh, J. I. Nurnberger Jr., H. J. Edenberg, E. S. Gershon, N. Schork, S. Bipolar Genome, N. D. Price, R. Gelinas, L. Hood, D. Craig, F. J. McMahon, J. R. Kelsoe, and J. C. Roach. Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proc Natl Acad Sci U S A*, 112(11):3576–3581, 2015.

[6] S. Anders, P. T. Pyl, and W. Huber. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[7] D. Aran, M. Sirota, and A. J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6, 2015.

[8] J. Asimit and E. Zeggini. Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics*, 44(1):293–308, 2010.

[9] D. Avramopoulos, V. K. Lasseter, M. D. Fallin, P. S. Wolyniec, J. A. McGrath, G. Nestadt, D. Valle, and A. E. Pulver. Stage II follow-up on a linkage scan for bipolar disorder in the Ashkenazim provides suggestive evidence for chromosome 12p and the GRIN2B gene. *Genetics in Medicine*, 9(11):745–751, 2007.

[10] J. A. Badner, D. Koller, T. Foroud, H. Edenberg, J. I. Nurnberger Jr., P. P. Zandi, V. L. Willour, F. J. McMahon, J. B. Potash, M. Hamshere, D. Grozeva, E. Green, G. Kirov, I. Jones, L. Jones, N. Craddock, D. Morris, R. Segurado, M. Gill, D. Sadovnick, R. Remick, P. Keck, J. Kelsoe, M. Ayub, A. MacLean, D. Blackwood, C. Y. Liu, E. S. Gershon, W. McMahon, G. J. Lyon, R. Robinson, J. Ross, and W. Byerley. Genome-wide linkage analysis of 972 bipolar pedigrees using single-nucleotide polymorphisms. *Mol Psychiatry*, 17(8):818–826, 2012.

[11] J. A. Bailey, D. M. Church, M. Ventura, M. Rocchi, and E. E. Eichler. Analysis of segmental duplications and genome assembly in the mouse. *Genome Research*, 14(5):789–801, 2004.

[12] R. S. Balaban, S. Nemoto, and T. Finkel. Mitochondria, oxidants, and aging. *Cell*, 120(4):483–495, 2005.

[13] R. V. Basaiawmoit and S. I. Rattan. Cellular stress and protein misfolding during aging. *Methods Mol Biol*, 648:107–117, 2010.

[14] A. Baud, V. Guryev, O. Hummel, M. Johannesson, J. Flint, R. Hermsen, P. Stridh, D. Graham, M. W. McBride, T. Foroud, S. Calderari, M. Diez, J. Ockinger, A. D. Beyeen, A. Gillett, N. Abdelmagid, A. O. Guerreiro-Cacais, M. Jagodic, J. Tuncel, U. Norin, E. Beattie, N. Huynh, W. H. Miller, D. L. Koller, I. Alam, S. Falak, M. Osborne-Pellegrin, E. Martinez-Membrives, T. Canete, G. Blazquez, E. Vicens-Costa, C. Mont-Cardona, S. Diaz-Moran, A. Tobena, D. Zelenika, K. Saar, G. Patone, A. Bauerfeind, M. T. Bihoreau, M. Heinig, Y. A. Lee, C. Rintisch, H. Schulz, D. A. Wheeler, K. C. Worley, D. M. Muzny, R. A. Gibbs, M. Lathrop, N. Lansu, P. Toonen, F. P. Ruzius, E. De Bruijn, H. Hauser, D. J. Adams, T. Keane, S. S. Atanur, T. J. Aitman, P. Flicek, T. Malinauskas, E. Y. Jones, D. Ekman, R. Lopez-Aumatell, A. F. Dominiczak, R. Holmdahl, T. Olsson, D. Gauguier, N. Hubner, A. Fernandez-Teruel, E. Cuppen, and R. Mott. Genomes and phenomes of a population of outbred rats and its progenitors. *Scientific Data*, 1, 2014.

[15] F. Benedetti and S. Dallaspezia. Melatonin, circadian rhythms, and the clock genes in bipolar disorder, 2009.

[16] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32:314–331, 1980.

[17] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic, 2017.

[18] Broad Institute. Picard, 2018.

[19] B. L. Browning and S. R. Browning. A fast, powerful method for detecting identity by descent. *Am J Hum Genet*, 88(2):173–182, 2011.

[20] C. Bycroft, C. Freeman, D. Petkova, G. Band, O. Delaneau, J. O. Connell, A. Cortes, and S. Welsh. Genome-wide genetic data on ˜ 500 , 000 UK Biobank participants. *bioRxiv*, 2017.

[21] N. Cai, T. B. Bigdeli, W. W. Kretzschmar, Y. Li, J. Liang, J. Hu, R. E. Peterson, S. Bacanu, B. T. Webb, B. Riley, Q. Li, J. Marchini, R. Mott, K. S. Kendler, and J. Flint. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Scientific Data*, 4, 2017.

[22] V. Calabrese, A. M. Stella, D. A. Butterfield, and G. Scapagnini. Redox regulation in neurodegeneration and longevity: role of the heme oxygenase and HSP70 systems in brain stress tolerance. *Antioxid Redox Signal*, 6(5):895–913, 2004.

[23] M. U. Carnes, R. R. Allingham, A. Ashley-Koch, and M. A. Hauser. Transcriptome analysis of adult and fetal trabecular meshwork, cornea, and ciliary body tissues by RNA sequencing. *Exp Eye Res*, 2016.

[24] C. Cavallotti, J. Feher, N. Pescosolido, and P. Sagnelli. Glycosaminoglycans in human trabecular meshwork: age-related changes. *Ophthalmic Res*, 36(4):211–217, 2004.

[25] F. Chen, X. Ding, Y. Feng, T. Seebeck, Y. Jiang, and G. D. Davis. Targeted activation of diverse CRISPR-Cas systems for mammalian genome editing via proximal CRISPR targeting. *Nature Communications*, 8, 2017.

[26] S. Y. Cherlyn, P. S. Woon, J. J. Liu, W. Y. Ong, G. C. Tsai, and K. Sim. Genetic association studies of glutamate, GABA and related genes in schizophrenia and bipolar disorder: a decade of advance. *Neurosci Biobehav Rev*, 34(6):958–977, 2010.

[27] V. Cheung and R. Spielman. The genetics of variation in gene expression. *Nature genetics*, 32 Suppl:522–525, 2002.

[28] M. Claussnitzer, S. N. Dankel, K.-H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I. S. Sousa, J. L. Beaudry, V. Puviindran, N. A. Abdennur, J. Liu, P.-A. Svensson, Y.-H. Hsu, D. J. Drucker, G. Mellgren, C.-C. Hui, H. Hauner, and M. Kellis. <i>FTO</i> Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*, 373(10):895–907, 2015.

[29] N. Craddock and L. Forty. Genetics of affective (mood) disorders. *Eur J Hum Genet*, 14(6):660–668, 2006.

[30] C. Cruceanu, A. Ambalavanan, D. Spiegelman, J. Gauthier, R. G. Lafreniere, P. A. Dion, M. Alda, G. Turecki, and G. A. Rouleau. Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. *Genome*, 56(10):634–640, 2013.

[31] D. Curtis. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *bioRxiv*, page 287136, 2018.

[32] J. P. de Magalhaes, J. Curado, and G. M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875–881, 2009.

[33] S. D. Detera-Wadleigh, J. A. Badner, W. H. Berrettini, T. Yoshikawa, L. R. Goldin, G. Turner, D. Y. Rollins, T. Moses, A. R. Sanders, J. D. Karkera, L. E. Esterling, J. Zeng, T. N. Ferraro, J. J. Guroff, D. Kazuba, M. E. Maxwell, J. I. Nurnberger, and E. S. Gershon. A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. *Proceedings of the National Academy of Sciences*, 96(10):5604–5609, 1999.

[34] S. J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, 36(19), 2008.

[35] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[36] ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[37] E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond, 2013.

[38] H. Ewald, T. Flint, T. A. Kruse, and O. Mors. A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24.3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13.3. *Molecular Psychiatry*, 7(7):734–744, 2002.

[39] K. K. H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2015.

[40] M. W. Feldman and S. Ramachandran. Missing compared to what? Revisiting heritability, genes and culture. *Philosophical Transactions of the Royal Society, B*, 373:20170064, 2018.

[41] M. A. Ferreira, M. C. O'Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, L. Jones, J. Fan, G. Kirov, R. H. Perlis, E. K. Green, J. W. Smoller,

D. Grozeva, J. Stone, I. Nikolov, K. Chambert, M. L. Hamshere, V. L. Nimgaonkar, V. Moskvina, M. E. Thase, S. Caesar, G. S. Sachs, J. Franklin, K. Gordon-Smith, K. G. Ardlie, S. B. Gabriel, C. Fraser, B. Blumenstiel, M. Defelice, G. Breen, M. Gill, D. W. Morris, A. Elkin, W. J. Muir, K. A. McGhee, R. Williamson, D. J. MacIntyre, A. W. MacLean, C. D. St, M. Robinson, M. Van Beck, A. C. Pereira, R. Kandaswamy, A. McQuillin, D. A. Collier, N. J. Bass, A. H. Young, J. Lawrence, I. N. Ferrier, A. Anjorin, A. Farmer, D. Curtis, E. M. Scolnick, P. McGuffin, M. J. Daly, A. P. Corvin, P. A. Holmans, D. H. Blackwood, H. M. Gurling, M. J. Owen, S. M. Purcell, P. Sklar, and N. Craddock. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet*, 40(9):1056–1058, 2008.

[42] H. B. Fraser, P. Khaitovich, J. B. Plotkin, S. Paabo, and M. B. Eisen. Aging and gene expression in the primate brain. *PLoS Biol*, 3(9):e274, 2005.

[43] L. G. Fritsche, W. Igl, J. N. C. Bailey, F. Grassmann, S. Sengupta, J. L. Bragg-Gresham, K. P. Burdon, S. J. Hebbring, C. Wen, M. Gorski, I. K. Kim, D. Cho, D. Zack, E. Souied, H. P. N. Scholl, E. Bala, K. E. Lee, D. J. Hunter, R. J. Sardell, P. Mitchell, J. E. Merriam, V. Cipriani, J. D. Hoffman, T. Schick, Y. T. E. Lechanteur, R. H. Guymer, M. P. Johnson, Y. Jiang, C. M. Stanton, G. H. S. Buitendijk, X. Zhan, A. M. Kwong, A. Boleda, M. Brooks, L. Gieser, R. Ratnapriya, K. E. Branham, J. R. Foerster, J. R. Heckenlively, M. I. Othman, B. J. Vote, H. H. Liang, E. Souzeau, I. L. McAllister, T. Isaacs, J. Hall, S. Lake, D. A. Mackey, I. J. Constable, J. E. Craig, T. E. Kitchner, Z. Yang, Z. Su, H. Luo, D. Chen, H. Ouyang, K. Flagg, D. Lin, G. Mao, H. Ferreyra, K. Stark, C. N. von Strachwitz, A. Wolf, C. Brandl, G. Rudolph, M. Olden, M. A. Morrison, D. J. Morgan, M. Schu, J. Ahn, G. Silvestri, E. E. Tsironi, K. H. Park, L. A. Farrer, A. Orlin, A. Brucker, M. Li, C. A. Curcio, S. Mohand-Saïd, J.-A. Sahel, I. Audo, M. Benchaboune, A. J. Cree, C. A. Rennie, S. V. Goverdhan, M. Grunin, S. Hagbi-Levi, P. Campochiaro, N. Katsanis, F. G. Holz, F. Blond, H. Blanché, J.-F. Deleuze, R. P. Igo, B. Truitt, N. S. Peachey, S. M. Meuer, C. E. Myers, E. L. Moore, R. Klein, M. A. Hauser, E. A. Postel, M. D. Courtenay, S. G. Schwartz, J. L. Kovach, W. K. Scott, G. Liew, A. G. Tan, B. Gopinath, J. C. Merriam, R. T. Smith, J. C. Khan, H. Shahid, A. T. Moore, J. A. McGrath, R. Laux, M. A. Brantley, A. Agarwal, L. Ersoy, A. Caramoy, T. Langmann, N. T. M. Saksens, E. K. de Jong, C. B. Hoyng, M. S. Cain, A. J. Richardson, T. M. Martin, J. Blangero, D. E. Weeks, B. Dhillon, C. M. van Duijn, K. F. Doheny, J. Romm, C. C. W. Klaver, C. Hayward, M. B. Gorin, M. L. Klein, P. N. Baird, A. I. den Hollander, S. Fauser, J. R. W. Yates, R. Allikmets, J. J. Wang, D. A. Schaumberg, B. E. K. Klein, S. A. Hagstrom, I. Chowers, A. J. Lotery, T. Léveillard, K. Zhang, M. H. Brilliant, A. W. Hewitt, A. Swaroop, E. Y. Chew, M. A. Pericak-Vance, M. DeAngelis, D. Stambolian, J. L. Haines, S. K. Iyengar, B. H. F. Weber, G. R. Abecasis, and I. M. Heid. A large genome-wide association study of age-related macu-

lar degeneration highlights contributions of rare and common variants. *Nature genetics*, 48(2):134–43, 2016.

[44] M. Gallagher and A. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics*, 102(5):717–730, 2018.

[45] B. Georgi, D. Craig, R. L. Kember, W. Liu, I. Lindquist, S. Nasser, C. Brown, J. A. Egeland, S. M. Paul, and M. Bucan. Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genet*, 10(3):e1004229, 2014.

[46] R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. Deramo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. L. Kraft, T. Nguyen, C. M. Pfannkoch, C. Sitter, G. G. Sutton, J. C. Venter, T. Woodage, D. Smith, H. M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. B. Weiss, D. M. Dunn, E. D. Green, R. W. Blakesley, G. G. Bouffard, P. J. de Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. C. Nierman, R. A. Gibbs, G. M. Weinstock, P. H. Havlak, R. Chen, K. J. Durbin, A. Egan, Y. Ren, X. Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, G. M. Weinstock, K. C. Worley, A. J. Cooney, R. A. Gibbs, L. M. D'Souza, K. Martin, J. Q. Wu, M. L. Gonzalez-Garay, A. R. Jackson, K. J. Kalafus, M. P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. A. Wheeler, Z. Zhang, J. A. Bailey, E. E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. J. Trask, J. M. Young, D. Smith, H. Huang, K. Fechtel, H. Wang, H. Xing, K. Weinstock, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. M. Albà, J. F. Abril, R. Guigo, A. Smit, I. Dubchak, E. M. Rubin, O. Couronne, A. Poliakov, N. Hübner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y. A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. J. Jacob, S. Bromberg, J. Gullings-Handley, M. I. Jensen-Seaman, A. E. Kwitek, J. Lazar, D. Pasko, P. J. Tonellato, S. Twigger, C. P. Ponting, J. M. Duarte, S. Rice, L. Goodstadt, S. A. Beatson, R. D. Emes, E. E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S. M. Searle, G. M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. A. Stone, J. C. Venter, B. A. Payseur, G. Bourque, C. López-Otín, X. S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. B. Yap, A. Caspi, S. D. G. Tesler, P. A. Pevzner, S. C. D. Haussler, K. M. Roskin, R. Baertsch, H. Clawson, T. S. Furey,

A. S. Hinrichs, D. Karolchik, W. J. Kent, K. R. Rosenbloom, H. Trumbower, M. Weirauch, D. N. Cooper, P. D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. R. Copley, M. S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, and F. Collins. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–520, 2004.

[47] A. Gillet-Markowska, H. Richard, G. Fischer, and I. Lafontaine. Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics*, 31(6):801–808, 2015.

[48] D. Glass, A. Viñuela, M. N. Davies, A. Ramasamy, L. Parts, D. Knowles, A. A. Brown, Å. K. Hedman, K. S. Small, A. Buil, E. Grundberg, A. C. Nica, P. Meglio, F. O. Nestle, M. Ryten, U. K. B. E. consortium The, T. consortium the Mu, R. Durbin, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, M. E. Weale, V. Bataille, and T. D. Spector. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biology*, 14(7):R75, 2013.

[49] M. Goel, R. G. Picciani, R. K. Lee, and S. K. Bhattacharya. Aqueous humor dynamics: a review. *Open Ophthalmol J*, 4:52–59, 2010.

[50] F. S. Goes, M. Pirooznia, J. S. Parla, M. Kramer, E. Ghiban, S. Mavruk, Y. C. Chen, E. T. Monson, V. L. Willour, R. Karchin, M. Flickinger, A. E. Locke, S. E. Levy, L. J. Scott, M. Boehnke, E. Stahl, J. L. Moran, C. M. Hultman, M. Landen, S. M. Purcell, P. Sklar, P. P. Zandi, W. R. McCombie, and J. B. Potash. Exome Sequencing of Familial Bipolar Disorder. *JAMA Psychiatry*, 73(6):590–597, 2016.

[51] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. On the immortality of television sets: ”Function” in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution*, 5(3):578–590, 2013.

[52] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale, 2003.

[53] S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, S. Consortium on the Genetics of, P. S. Group, V. L. Nimgaonkar, R. C. Go, R. M. Savage, N. R. Swerdlow, R. E. Gur, D. L. Braff, M. C. King, and J. M. McClellan. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, 154(3):518–529, 2013.

[54] D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M. O. Pollard, A. Choudhury, G. R. Ritchie, Y. Xue, J. Asimit, R. N. Nsubuga, E. H. Young, C. Pomilla, K. Kivinen, K. Rockett, A. Kamali, A. P. Doumatey, G. Asiki, J. Seeley, F. Sisay-Joof,

M. Jallow, S. Tollman, E. Mekonnen, R. Ekong, T. Oljira, N. Bradman, K. Bojang, M. Ramsay, A. Adeyemo, E. Bekele, A. Motala, S. A. Norris, F. Pirie, P. Kaleebu, D. Kwiatkowski, C. Tyler-Smith, C. Rotimi, E. Zeggini, and M. S. Sandhu. The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534):327–332, 2015.

[55] V. Guryev, K. Saar, T. Adamovic, M. Verheul, S. A. A. C. Van Heesch, S. Cook, M. Pravenec, T. Aitman, H. Jacob, J. D. Shull, N. Hubner, and E. Cuppen. Distribution and functional impact of DNA copy number variation in the rat. *Nature Genetics*, 40:538–545, 2008.

[56] HAGR. GenAge.

[57] R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, and S. A. Mccarroll. Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3):296–303, 2015.

[58] C. Hansen and K. Spuhler. Development of the National Institutes of Health Genetically Heterogeneous Rat Stock. *Alcoholism: Clinical and Experimental Research*, 8(5):477–479, 1984.

[59] A. G. Harvey. Sleep and circadian rhythms in bipolar disorder: seeking synchrony, harmony, and regulation. *Am J Psychiatry*, 165(7):820–829, 2008.

[60] E. Heitzer, P. Ulz, J. Belic, S. Gutschi, F. Quehenberger, K. Fischereder, T. Benezeder, M. Auer, C. Pischler, S. Mannweiler, M. Pichler, F. Eisner, M. Haeusler, S. Riethdorf, K. Pantel, H. Samonigg, G. Hoefler, H. Augustin, J. B. Geigl, and M. R. Speicher. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Medicine*, 5(4), 2013.

[61] R. Hermsen, J. de Ligt, W. Spee, F. Blokzijl, S. Schäfer, E. Adami, S. Boymans, S. Flink, R. van Boxtel, R. H. van der Weide, T. Aitman, N. Hübner, M. Simonis, B. Tabakoff, V. Guryev, and E. Cuppen. Genomic landscape of rat strain and substrain variation. *BMC Genomics*, 16, 2015.

[62] L. Hirschfeld and H. Hirschfeld. Serological differences between the blood of different races. The result of researches on the Macedonian front. *Lancet*, 194(5016):675–679, 1919.

[63] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn. A comprehensive review of genetic association studies. *Genet Med*, 4(2):45–61, 2002.

[64] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. C. Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19(7):1270–1278, 2009.

[65] S. Horvath. DNA methylation age of human tissues and cell types. *Genome Biol*, 14(10):R115, 2013.

[66] W. Huang da, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[67] G. M. Hughes, E. M. Boston, J. A. Finarelli, W. J. Murphy, D. G. Higgins, and E. C. Teeling. The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Molecular Biology and Evolution*, (March 2018):1–49, 2018.

[68] P. M. Iannaccone and H. J. Jacob. Rats! *Dis Model Mech*, 2(5-6):206–210, 2009.

[69] T. Indian and G. Variation. The Indian Genome Variation database (IGVdb): a project overview. *Human genetics*, 118(1):1–11, 2005.

[70] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[71] Y. Jiang, Y. Wang, and M. Brudno. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, 28(20):2576–2583, 2012.

[72] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–354, 2010.

[73] R. J. Klein. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720):385–389, 2005.

[74] L. G. KOCH and S. L. BRITTON. Artificial selection for intrinsic aerobic endurance running capacity in rats. *Physiological Genomics*, 5(1):45–52, 2001.

[75] M. E. Koran, T. a. Thornton-Wells, N. Jahanshad, D. C. Glahn, P. M. Thompson, J. Blangero, T. E. Nichols, P. Kochunov, and B. a. Landman. Impact of family structure and common environment on heritability estimation for neuroimaging genetics studies using Sequential Oligogenic Linkage Analysis Routines. *Journal of medical imaging (Bellingham, Wash.)*, 1(1):14005, 2014.

[76] J. O. Korbel, A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein. PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2), 2009.

[77] D. F. Kripke, D. J. Mullaney, M. Atkinson, and S. Wolf. Circadian rhythm disorders in manic-depressives. *Biol Psychiatry*, 13(3):335–351, 1978.

[78] A. Kumar, J. R. Gibbs, A. Beilina, A. Dillman, R. Kumaran, D. Trabzuni, M. Ryten, R. Walker, C. Smith, B. J. Traynor, J. Hardy, A. B. Singleton, and M. R. Cookson. Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiol Aging*, 34(4):1199–1209, 2013.

[79] E. Lander and R. Weinberg. Genomics: journey to the center of biology. *Science*, 287(5459):1777–1782, 2000.

[80] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[81] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, 15(6), 2014.

[82] C. K. Lee, R. Weindruch, and T. A. Prolla. Gene-expression profile of the ageing brain in mice. *Nat Genet*, 25(3):294–297, 2000.

[83] N. J. Leeper, J. Myers, M. Zhou, K. T. Nead, A. Syed, Y. Kojima, R. D. Caceres, and J. P. Cooke. Exercise capacity is the strongest predictor of mortality in patients with peripheral arterial disease. *Journal of Vascular Surgery*, 57(3):728–733, 2013.

[84] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[85] J. Z. Li, B. G. Bunney, F. Meng, M. H. Hagenauer, D. M. Walsh, M. P. Vawter, S. J. Evans, P. V. Choudary, P. Cartagena, J. D. Barchas, A. F. Schatzberg, E. G. Jones, R. M. Myers, S. J. Watson Jr., H. Akil, and W. E. Bunney. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proc Natl Acad Sci U S A*, 110(24):9950–9955, 2013.

[86] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9:271, 2008.

[87] A. Llobet, X. Gasull, and A. Gual. Understanding Trabecular Meshwork Physiology: A Key to the Control of Intraocular Pressure? *Physiology*, 18(5):205–209, 2003.

[88] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, Daniel MacArthur, M. Kellis, A. Thomson, T. Young,

E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A Roger Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project, 2013.

[89] J. Lund, P. Tedesco, K. Duke, J. Wang, S. K. Kim, and T. E. Johnson. Transcriptional profile of aging in C. elegans. *Curr Biol*, 12(18):1566–1573, 2002.

[90] G. Lunter and M. Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011.

[91] M. Manczak, Y. Jung, B. S. Park, D. Partovi, and P. H. Reddy. Time-course of mitochondrial gene expressions in mice brains: implications for mitochondrial dysfunction, oxidative damage, and cytochrome c in aging. *J Neurochem*, 92(3):494–504, 2005.

[92] A. R. Martin, C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, S. Gravel, M. J. Daly, C. D. Bustamante, and E. E. Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635–649, 2017.

[93] C. A. McClung. Role for the Clock gene in bipolar disorder. *Cold Spring Harb Symp Quant Biol*, 72:637–644, 2007.

[94] C. L. McGrath, S. J. Glatt, P. Sklar, H. Le-Niculescu, R. Kuczenski, A. E. Doyle, J. Biederman, E. Mick, S. V. Faraone, A. B. Niculescu, and M. T. Tsuang. Evidence for genetic association of RORB with bipolar disorder. *BMC Psychiatry*, 9, 2009.

[95] P. McGuffin, F. Rijsdijk, M. Andrew, P. Sham, R. Katz, and A. Cardno. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Archives of general psychiatry*, 60(5):497–502, 2003.

[96] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, 2010.

[97] M. B. McQueen, B. Devlin, S. V. Faraone, V. L. Nimgaonkar, P. Sklar, J. W. Smoller, R. Abou Jamra, M. Albus, S. A. Bacanu, M. Baron, T. B. Barrett,

W. Berrettini, D. Blacker, W. Byerley, S. Cichon, W. Coryell, N. Craddock, M. J. Daly, J. R. Depaulo, H. J. Edenberg, T. Foroud, M. Gill, T. C. Gilliam, M. Hamshere, I. Jones, L. Jones, S. H. Juo, J. R. Kelsoe, D. Lambert, C. Lange, B. Lerer, J. Liu, W. Maier, J. D. Mackinnon, M. G. McInnis, F. J. McMahon, D. L. Murphy, M. M. Nothen, J. I. Nurnberger, C. N. Pato, M. T. Pato, J. B. Potash, P. Propping, A. E. Pulver, J. P. Rice, M. Rietschel, W. Scheftner, J. Schumacher, R. Segurado, K. Van Steen, W. Xie, P. P. Zandi, and N. M. Laird. Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *Am J Hum Genet*, 77(4):582–595, 2005.

[98] J. a. Miller, S.-L. Ding, S. M. Sunkin, K. a. Smith, L. Ng, A. Szafer, A. Ebbert, Z. L. Riley, J. J. Royall, K. Aiona, J. M. Arnold, C. Bennet, D. Bertagnolli, K. Brouner, S. Butler, S. Caldejon, A. Carey, C. Cuhaciyan, R. a. Dalley, N. Dee, T. a. Dolbeare, B. a. C. Facer, D. Feng, T. P. Fliss, G. Gee, J. Goldy, L. Gourley, B. W. Gregor, G. Gu, R. E. Howard, J. M. Jochim, C. L. Kuan, C. Lau, C.-K. Lee, F. Lee, T. a. Lemon, P. Lesnar, B. McMurray, N. Mastan, N. Mosqueda, T. Naluai-Cecchini, N.-K. Ngo, J. Nyhus, A. Oldre, E. Olson, J. Parente, P. D. Parker, S. E. Parry, A. Stevens, M. Pletikos, M. Reding, K. Roll, D. Sandman, M. Sarreal, S. Shapouri, N. V. Shapovalova, E. H. Shen, N. Sjoquist, C. R. Slaughterbeck, M. Smith, A. J. Sodt, D. Williams, L. Zöllei, B. Fischl, M. B. Gerstein, D. H. Geschwind, I. a. Glass, M. J. Hawrylycz, R. F. Hevner, H. Huang, A. R. Jones, J. a. Knowles, P. Levitt, J. W. Phillips, N. Sestan, P. Wohnoutka, C. Dang, A. Bernard, J. G. Hohmann, and E. S. Lein. Transcriptional landscape of the prenatal human brain. *Nature*, 508:199–206, 2014.

[99] J. T. Morgan, V. K. Raghunathan, Y. R. Chang, C. J. Murphy, and P. Russell. The intrinsic stiffness of human trabecular meshwork cells increases with senescence. *Oncotarget*, 6(17):15362–15374, 2015.

[100] S. Mukherjee. *The Gene An Intimate History.* Penguin Books India, 1 edition, 2016.

[101] B. M. Neale and P. Sklar. Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but also suggests new directions for molecular interrogation. *Curr Opin Neurobiol*, 30:131–138, 2015.

[102] A. C. Need, J. P. McEvoy, M. Gennarelli, E. L. Heinzen, D. Ge, J. M. Maia, K. V. Shianna, M. He, E. T. Cirulli, C. E. Gumbs, Q. Zhao, C. R. Campbell, L. Hong, P. Rosenquist, A. Putkonen, T. Hallikainen, E. Repo-Tiihonen, J. Tiihonen, D. L. Levy, H. Y. Meltzer, and D. B. Goldstein. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet*, 91(2):303–312, 2012.

[103] P. C. Ng and S. Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, 2003.

[104] C. M. Nievergelt, D. F. Kripke, T. B. Barrett, E. Burg, R. A. Remick, A. D. Sadovnick, S. L. McElroy, P. E. Keck, N. J. Schork, and J. R. Kelsoe. Suggestive evidence for association of the circadian genes PERIOD3 and ARNTL with bipolar disorder. *American Journal of Medical Genetics - Neuropsychiatric Genetics*, 141 B(3):234–241, 2006.

[105] R. E. Nisbett, J. Aronson, C. Blair, W. Dickens, J. Flynn, D. F. Halpern, and E. Turkheimer. Intelligence: New Findings and Theoretical Developments. *American Psychologist*, 67(2):130–159, 2012.

[106] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[107] M. J. Peters, R. Joehanes, L. C. Pilling, C. Schurmann, K. N. Conneely, J. Powell, E. Reinmaa, G. L. Sutphin, A. Zhernakova, K. Schramm, Y. A. Wilson, S. Kobes, T. Tukiainen, M. A. Nalls, D. G. Hernandez, M. R. Cookson, R. J. Gibbs, J. Hardy, A. Ramasamy, A. B. Zonderman, A. Dillman, B. Traynor, C. Smith, D. L. Longo, D. Trabzuni, J. Troncoso, M. van der Brug, M. E. Weale, R. O'Brien, R. Johnson, R. Walker, R. H. Zielke, S. Arepalli, M. Ryten, A. B. Singleton, Y. F. Ramos, H. H. H. Göring, M. Fornage, Y. Liu, S. A. Gharib, B. E. Stranger, P. L. De Jager, A. Aviv, D. Levy, J. M. Murabito, P. J. Munson, T. Huan, A. Hofman, A. G. Uitterlinden, F. Rivadeneira, J. van Rooij, L. Stolk, L. Broer, M. M. P. J. Verbiest, M. Jhamai, P. Arp, A. Metspalu, L. Tserel, L. Milani, N. J. Samani, P. Peterson, S. Kasela, V. Codd, A. Peters, C. K. Ward-Caviness, C. Herder, M. Waldenberger, M. Roden, P. Singmann, S. Zeilinger, T. Illig, G. Homuth, H.-J. Grabe, H. Völzke, L. Steil, T. Kocher, A. Murray, D. Melzer, H. Yaghootkar, S. Bandinelli, E. K. Moses, J. W. Kent, J. E. Curran, M. P. Johnson, S. Williams-Blangero, H.-J. Westra, A. F. McRae, J. A. Smith, S. L. R. Kardia, I. Hovatta, M. Perola, S. Ripatti, V. Salomaa, A. K. Henders, N. G. Martin, A. K. Smith, D. Mehta, E. B. Binder, K. M. Nylocks, E. M. Kennedy, T. Klengel, J. Ding, A. M. Suchy-Dicey, D. A. Enquobahrie, J. Brody, J. I. Rotter, Y.-D. I. Chen, J. Houwing-Duistermaat, M. Kloppenburg, P. E. Slagboom, Q. Helmer, W. den Hollander, S. Bean, T. Raj, N. Bakhshi, Q. P. Wang, L. J. Oyston, B. M. Psaty, R. P. Tracy, G. W. Montgomery, S. T. Turner, J. Blangero, I. Meulenbelt, K. J. Ressler, J. Yang, L. Franke, J. Kettunen, P. M. Visscher, G. G. Neely, R. Korstanje, R. L. Hanson, H. Prokisch, L. Ferrucci, T. Esko, A. Teumer, J. B. J. van Meurs, and A. D. Johnson. The transcriptional landscape of age in human peripheral blood. *Nature Communications*, 6:8570, 2015.

[108] B. S. Pickard, E. J. Hollox, M. P. Malloy, D. J. Porteous, D. H. Blackwood, J. A. Armour, and W. J. Muir. A 4q35.2 subtelomeric deletion identified in

a screen of patients with co-morbid psychiatric illness and mental retardation. *BMC Medical Genetics*, 5, 2004.

[109] S. D. Pletcher, S. J. Macdonald, R. Marguerie, U. Certa, S. C. Stearns, D. B. Goldstein, and L. Partridge. Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. *Curr Biol*, 12(9):712–723, 2002.

[110] M. D. Preston and F. Dudbridge. Utilising family-based designs for detecting rare variant disease associations. *Ann Hum Genet*, 78(2):129–140, 2014.

[111] Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet*, 43(10):977–983, 2011.

[112] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, 2007.

[113] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kahler, L. Duncan, E. Stahl, G. Genovese, E. Fernandez, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, and P. Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190, 2014.

[114] S. Ramdas. No Title.

[115] A. R. Rao, M. Yourshaw, B. Christensen, S. F. Nelson, and B. Kerner. Rare deleterious mutations are associated with disease in bipolar disorder families. *Mol Psychiatry*, 22(7):1009–1014, 2017.

[116] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease, 2001.

[117] Y.-y. Ren, L. Koch, S. Britton, N. Qi, M. K. Treutelaar, C. F. Burant, and J. Z. Li. High-density SNP array and genome sequencing reveal signatures of selection in a divergent selection rat model for aerobic running capacity. *bioRxiv*, 2016.

[118] Y.-y. Ren, L. G. Koch, S. L. Britton, N. R. Qi, M. K. Treutelaar, C. F. Burant, and J. Z. Li. Selection-, age-, and exercise-dependence of skeletal muscle gene expression patterns in a rat model of metabolic fitness. *Physiological Genomics*, 48(11):816–825, 2016.

[119] S. Ripke, C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kahler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, Y. Kim, S. H. Lee, P. K. Magnusson, N. Sanchez, E. A. Stahl, S. Williams, N. R. Wray, K. Xia, F. Bettella, A. D. Borglum, B. K. Bulik-Sullivan, P. Cormican, N. Craddock,

C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M. L. Hamshere, P. Holmans, D. M. Hougaard, K. S. Kendler, K. Lin, D. W. Morris, O. Mors, P. B. Mortensen, B. M. Neale, F. A. O'Neill, M. J. Owen, M. P. Milovancevic, D. Posthuma, J. Powell, A. L. Richards, B. P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A. B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari, S. Tosato, M. Verhage, J. T. Walters, C. Multicenter Genetic Studies of Schizophrenia, D. F. Levinson, P. V. Gejman, K. S. Kendler, C. Laurent, B. J. Mowry, M. C. O'Donovan, M. J. Owen, A. E. Pulver, B. P. Riley, S. G. Schwab, D. B. Wildenauer, F. Dudbridge, P. Holmans, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F. B. Lerer, K. Y. Liang, W. Maier, J. Mallet, D. A. Nertney, G. Nestadt, N. Norton, F. A. O'Neill, G. N. Papadimitriou, R. Ribble, A. R. Sanders, J. M. Silverman, D. Walsh, N. M. Williams, B. Wormley, C. Psychosis Endophenotypes International, M. J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, J. Hall, C. Iyegbe, A. Jablensky, R. S. Kahn, L. Kalaydjieva, S. Lawrie, C. M. Lewis, K. Lin, D. H. Linszen, I. Mata, A. McIntosh, R. M. Murray, R. A. Ophoff, J. Powell, D. Rujescu, J. Van Os, M. Walshe, M. Weisbrod, D. Wiersma, C. Wellcome Trust Case Control, P. Donnelly, I. Barroso, J. M. Blackwell, E. Bramon, M. A. Brown, J. P. Casas, A. P. Corvin, P. Deloukas, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, C. C. Spencer, G. Band, C. Bellenguez, C. Freeman, G. Hellenthal, E. Giannoulatou, M. Pirinen, R. D. Pearson, A. Strange, Z. Su, D. Vukcevic, P. Donnelly, C. Langford, S. E. Hunt, S. Edkins, R. Gwilliam, H. Blackburn, S. J. Bumpstead, S. Dronov, M. Gillman, E. Gray, N. Hammond, A. Jayakumar, O. T. McCann, J. Liddle, S. C. Potter, R. Ravindrarajah, M. Ricketts, A. Tashakkori-Ghanbaria, M. J. Waller, P. Weston, S. Widaa, P. Whittaker, I. Barroso, P. Deloukas, C. G. Mathew, J. M. Blackwell, M. A. Brown, A. P. Corvin, M. I. McCarthy, C. C. Spencer, E. Bramon, A. P. Corvin, M. C. O'Donovan, K. Stefansson, E. Scolnick, S. Purcell, S. A. McCarroll, P. Sklar, C. M. Hultman, and P. F. Sullivan. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*, 45(10):1150–1159, 2013.

[120] N. Risch and K. Merikangas. The Future of Genetic Studies of Complex Human Diseases. *Science*, 273(5281):1516–1517, 1996.

[121] G. E. J. Rodwell, R. Sonu, J. M. Zahn, J. Lund, J. Wilhelmy, L. Wang, W. Xiao, M. Mindrinos, E. Crane, E. Segal, B. D. Myers, J. D. Brooks, R. W. Davis, J. Higgins, A. B. Owen, and S. K. Kim. A Transcriptional Profile of Aging in the Human Kidney. *PLoS Biology*, 2(12):e427, 2004.

[122] A. Rutherford. *A Brief History of History Who Ever Lived.* Orion Publishing Group, 2016.

[123] M. A. Sartor, G. D. Leikauf, and M. Medvedovic. LRpath: a logistic regression

approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

[124] R. Segurado, S. D. Detera-Wadleigh, D. F. Levinson, C. M. Lewis, M. Gill, J. I. Nurnberger Jr., N. Craddock, J. R. DePaulo, M. Baron, E. S. Gershon, J. Ekholm, S. Cichon, G. Turecki, S. Claes, J. R. Kelsoe, P. R. Schofield, R. F. Badenhop, J. Morissette, H. Coon, D. Blackwood, L. A. McInnes, T. Foroud, H. J. Edenberg, T. Reich, J. P. Rice, A. Goate, M. G. McInnis, F. J. McMahon, J. A. Badner, L. R. Goldin, P. Bennett, V. L. Willour, P. P. Zandi, J. Liu, C. Gilliam, S. H. Juo, W. H. Berrettini, T. Yoshikawa, L. Peltonen, J. Lonnqvist, M. M. Nothen, J. Schumacher, C. Windemuth, M. Rietschel, P. Propping, W. Maier, M. Alda, P. Grof, G. A. Rouleau, J. Del-Favero, C. Van Broeckhoven, J. Mendlewicz, R. Adolfsson, M. A. Spence, H. Luebbert, L. J. Adams, J. A. Donald, P. B. Mitchell, N. Barden, E. Shink, W. Byerley, W. Muir, P. M. Visscher, S. Macgregor, H. Gurling, G. Kalsi, A. McQuillin, M. A. Escamilla, V. I. Reus, P. Leon, N. B. Freimer, H. Ewald, T. A. Kruse, O. Mors, U. Radhakrishna, J. L. Blouin, S. E. Antonarakis, and N. Akarsu. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *Am J Hum Genet*, 73(1):49–62, 2003.

[125] A. Serretti, F. Benedetti, L. Mandelli, C. Lorenzi, A. Pirovano, C. Colombo, and E. Smeraldi. Genetic dissection of psychopathological symptoms: insomnia in mood disorders and CLOCK gene polymorphism. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 121B:35–38, 2003.

[126] A. A. Shabalin. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[127] R. A. Shih, P. L. Belmonte, and P. P. Zandi. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders, 2004.

[128] E. Shink, J. Morissette, R. Sherrington, and N. Barden. A genome-wide scan points to a susceptibility locus for bipolar disorder on chromosome 12. *Molecular Psychiatry*, 10(6):545–552, 2005.

[129] G. Shinozaki and J. B. Potash. New developments in the genetics of bipolar disorder. *Curr Psychiatry Rep*, 16(11):493, 2014.

[130] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk. BioMart–biological queries made easy. *BMC Genomics*, 10:22, 2009.

[131] G. K. Smyth, J. Michaud, and H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

[132] L. C. Solberg Woods, C. Stelloh, K. R. Regner, T. Schwabe, J. Eisenhauer, and M. R. Garrett. Heterogeneous stock rats: a new model to study the genetics of renal phenotypes. *American journal of physiology. Renal physiology*, 298(6):F1484–91, 2010.

[133] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012.

[134] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V. E. Fraifeld, and J. P. de Magalhaes. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res*, 41(Database issue):D1027–33, 2013.

[135] A. Trifunovic and N. G. Larsson. Mitochondrial dysfunction as a cause of ageing. *J Intern Med*, 263(2):167–178, 2008.

[136] A. Varshney, L. J. Scott, R. P. Welch, M. R. Erdos, P. S. Chines, N. Narisu, R. D. Albanus, P. Orchard, B. N. Wolford, R. Kursawe, S. Vadlamudi, M. E. Cannon, J. P. Didion, J. Hensley, A. Kirilusha, L. L. Bonnycastle, D. L. Taylor, R. Watanabe, K. L. Mohlke, M. Boehnke, F. S. Collins, S. C. J. Parker, and M. L. Stitzel. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proceedings of the National Academy of Sciences*, 114(9):2301–2306, 2017.

[137] P. M. Visscher. Human complex trait genetics in the 21st century. *Genetics*, 202(2):377–379, 2016.

[138] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery, 2012.

[139] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, J. Zhang, Y. Guo, B. Feng, H. Li, Y. Lu, X. Fang, H. Liang, Z. Du, D. Li, Y. Zhao, Y. Hu, Z. Yang, H. Zheng, I. Hellmann, M. Inouye, J. Pool, X. Yi, J. Zhao, J. Duan, Y. Zhou, J. Qin, L. Ma, G. Li, Z. Yang, G. Zhang, B. Yang, C. Yu, F. Liang, W. Li, S. Li, D. Li, P. Ni, J. Ruan, Q. Li, H. Zhu, D. Liu, Z. Lu, N. Li, G. Guo, J. Zhang, J. Ye, L. Fang, Q. Hao, Q. Chen, Y. Liang, Y. Su, A. San, C. Ping, S. Yang, F. Chen, L. Li, K. Zhou, H. Zheng, Y. Ren, L. Yang, Y. Gao, G. Yang, Z. Li, X. Feng, K. Kristiansen, G. K. S. Wong, R. Nielsen, R. Durbin, L. Bolund, X. Zhang, S. Li, H. Yang, and J. Wang. The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, 2008.

[140] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[141] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(SUPPL. 2), 2010.

[142] M. M. Weissman, R. C. Bland, G. J. Canino, C. Faravelli, S. Greenwald, H. G. Hwu, P. R. Joyce, E. G. Karam, C. K. Lee, J. Lellouch, J. P. Lepine, S. C. Newman, M. Rubio-Stipec, J. E. Wells, P. J. Wickramaratne, H.-U. Wittchen, and E. K. Yeh. Cross-National Epidemiology of Major Depression and Bipolar Disorder. *Journal of the American Medical Association*, 276(4):293–299, 1996.

[143] A. J. Willsey, S. J. Sanders, M. Li, S. Dong, A. T. Tebbenkamp, R. A. Muhle, S. K. Reilly, L. Lin, S. Fertuzinhos, J. A. Miller, M. T. Murtha, C. Bichsel, W. Niu, J. Cotney, A. G. Ercan-Sencicek, J. Gockley, A. R. Gupta, W. Han, X. He, E. J. Hoffman, L. Klei, J. Lei, W. Liu, L. Liu, C. Lu, X. Xu, Y. Zhu, S. M. Mane, E. S. Lein, L. Wei, J. P. Noonan, K. Roeder, B. Devlin, N. Sestan, and M. W. State. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, 155(5):997–1007, 2013.

[144] B. Xu, J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet*, 43(9):864–868, 2011.

[145] J. Yang, T. Huang, F. Petralia, Q. Long, B. Zhang, C. Argmann, Y. Zhao, C. V. Mobbs, E. E. Schadt, J. Zhu, Z. Tu, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, W. Winckler, J. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shablin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. Goldman, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, L. Lin, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, C. Choi, E. Karasik, K. Ramsey, M. T. Moser, B. A. Foster, B. M. Gillard, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. Jewel, P. Branton, L. H. Sobin, L. Qi, P. Hariharan, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, B. E. Robles, M. Basile, D. C. Mash, S. Volpi, J. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, N. C. I. L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J.

Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, and N. C. Lockhart. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific Reports*, 5:15145, 2015.

[146] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, 2011.

[147] J. M. Zahn, S. Poosala, A. B. Owen, D. K. Ingram, A. Lustig, A. Carter, A. T. Weeraratna, D. D. Taub, M. Gorospe, K. Mazan-Mamczarz, E. G. Lakatta, K. R. Boheler, X. Xu, M. P. Mattson, G. Falco, M. S. Ko, D. Schlessinger, J. Firman, S. K. Kummerfeld, W. H. Wood 3rd, A. B. Zonderman, S. K. Kim, and K. G. Becker. AGEMAP: a gene expression database for aging in mice. *PLoS Genet*, 3(11):e201, 2007.

[148] J. M. Zahn, S. Poosala, A. B. Owen, A. B. Zonderman, and S. K. Kim. AGEMAP.

[149] J. M. Zahn, R. Sonu, H. Vogel, E. Crane, K. Mazan-Mamczarz, R. Rabkin, R. W. Davis, K. G. Becker, A. B. Owen, and S. K. Kim. Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet*, 2(7):e115, 2006.