12-2015

Functional Analysis of Genomic Variation and Impact on Molecular and Higher Order Phenotypes

Ashutosh Kumar Pandey
University of Tennessee Health Science Center

Recommended Citation

Pandey, Ashutosh Kumar , "Functional Analysis of Genomic Variation and Impact on Molecular and Higher Order Phenotypes" (2015). *Theses and Dissertations (ETD)*. Paper 359. http://dx.doi.org/10.21007/etd.cghs.2015.0237.

Functional Analysis of Genomic Variation and Impact on Molecular and Higher Order Phenotypes

Document Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Program
Biomedical Sciences

Track
Genetics, Functional Genomics, and Proteomics

Research Advisor
Robert W. Williams, Ph.D.

Committee
Hao Chen, Ph.D. Eldon E. Geisert, Ph.D. Ramin Homayouni, Ph.D. David R. Nelson, Ph.D.

DOI
10.21007/etd.cghs.2015.0237

Comments
Six month embargo expired June 2016

Functional Analysis of Genomic Variation and Impact on Molecular and Higher Order Phenotypes

A Dissertation
Presented for
The Graduate Studies Council
The University of Tennessee
Health Science Center

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy
From The University of Tennessee

By
Ashutosh Kumar Pandey
December 2015

DEDICATION

With great pride and affection, I dedicate this dissertation to my father for raising me single-handedly and sacrificing so much for my education.

ACKNOWLEDGEMENTS

I would like to express my deep appreciation and gratitude to my advisor, Dr. Robert W Williams, for providing me with the opportunity to perform doctoral research in his laboratory. My graduate training in his laboratory has been a great learning experience for me. He always encouraged me to develop independent thinking and research skills. His guidance, mentorship, and support throughout this research journey have been invaluable.

I would like to express sincere thanks to my dissertation committee members, Drs. Hao Chen, Eldon E Geisert, Ramin Homayouni, and David R Nelson, for their valuable suggestions and their precious time. I am also thankful to Drs. Donald B. Thomason, Patrick Ryan, and Rennolds Ostrom for all their support and help from the start to the end of my graduate study journey at UTHSC.

I extend my gratitude to Dr. Megan Mulligan, Dr. Xusheng Wang, Dr. Lu Lu, Lei Yan, Zachary Sloan, and all the other past and present members of the Williams' Lab, for being supportive and helpful during my graduate training.

I acknowledge all the funding sources that have sponsored my graduate studies (INIA grants U01AA016662, U01AA13499, U24AA13513, and U01AA014425; NIAAA grants U01AA016662, U01AA013499, U24AA013513, and U01AA014425; and the UTHSC Center for Integrative and Translational Genomics).

I am forever indebted to my parents for everything. I am grateful to my uncle, aunt, brother, and sister-in-law for their endless support and blessings in my pursuit of science. I am also grateful to my friend, Dr. Vinay Jain for his continuous moral support during all my graduate years.

ABSTRACT

Reverse genetics methods, particularly the production of gene knockouts and knockins, have revolutionized the understanding of gene function. High throughput sequencing now makes it practical to exploit reverse genetics to simultaneously study functions of thousands of normal sequence variants and spontaneous mutations that segregate in intercross and backcross progeny generated by mating completely sequenced parental lines. To evaluate this new reverse genetic method we resequenced the genome of one of the oldest inbred strains of mice—DBA/2J—the father of the large family of BXD recombinant inbred strains. We analyzed ~100X whole-genome sequence data for the DBA/2J strain, relative to C57BL/6J, the reference strain for all mouse genomics and the mother of the BXD family. We generated the most detailed picture of molecular variation between the two mouse strains to date and identified 5.4 million sequence polymorphisms, including, 4.46 million single nucleotide polymorphisms (SNPs), 0.94 million insertions/deletions (indels), and 20,000 structural variants. We systematically scanned massive databases of molecular phenotypes and ~4,000 classical phenotypes to detect linked functional consequences of sequence variants. In majority of cases we successfully recovered known genotype-to-phenotype associations and in several cases we linked sequence variants to novel phenotypes (*Ahr*, *Fh1, Entpd2,* and *Col6a5*). However, our most striking and consistent finding is that apparently deleterious homozygous SNPs, indels, and structural variants have undetectable or very modest additive effects on phenotypes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASE	Allele-specific expression
BMD	Bone mineral density
CNV	Copy number variant
EMR	Electronic medical record
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FPR	False positive rate
GERP	Genomic evolutionary rate profiling
GWAS	Genome-wide association study
HSR	Heat shock response
LOD	Logarithm of odds
LRS	Likelihood ratio statistic
MGP	Mouse genome project
NMD	Nonsense-mediated decay
PheWAS	Phenome-wide association study
QTL	Quantitative trait loci
RI	Recombinant inbred
TCA	Tricarboxylic acid
UPR	Unfolded protein response

CHAPTER 1. INTRODUCTION

INTRODUCTION TO GENOMIC MEDICINE

Genetic variation modulates virtually all aspects of phenotypic variation, including physiological and behavioral differences, susceptibility to diseases, and differences in drug responses among individuals within and between populations [1-5]. Identification of the sets of causal genetic variants is fundamental to a better understanding of molecular mechanisms underlying the heritable fraction of phenotypic variation. This, in turn, is important for developing more rational, mechanistic, and individualized strategies for therapeutic intervention.

The last decade has seen a significant shift from family-based linkage studies to genome-wide association studies [6-8]. The advent of high-throughput genotyping technology has enabled fast and accurate detection of genotypes for large numbers of known markers in large human cohorts. This, along with development of advanced statistical approaches (and ample funding), has propelled a series of systematic gene-to-disease association studies. To date, genetic association studies have reliably linked thousands of sequence variants (mainly SNPs) to a wide-variety of Mendelian and common complex diseases [9,10].

The dramatic reduction in the cost of high-throughput sequencing and computational resource required to analyze the data, has also enabled large scale sequencing efforts with the aim of comprehensive profiling genetic and transcriptomic differences between case and control cohorts [11-13]. The acquisition, integration and interpretation of multi-omic data including genomics and its derivatives—transcriptomics, proteomics and metabolomics—have been successful in the discovery of biomarkers for diagnostic, preventative and therapeutic purposes. These rapid developments are facilitating the clinical adoption of genomic medicine—the customization of treatment regimens and pharmaceuticals based on an individual's genetic profile [14-16]. Considering the rate at which sequencing technology is outpacing Moore's Law [17], characterizing genomic variation from personal genome sequencing will be as affordable as an X-ray scan. Phase I of genomic medicine is already having an impact on healthcare by offering genome-based diagnostic approaches for the prediction of disease risks, prediction of drug response, accurate molecular classification of disease, and early detection of diseases.

IDENTIFICATION OF DISEASE RISK FACTORS—THEN AND NOW

Genetic mapping approaches to identify disease risk factors have evolved largely during the last three decades [18,19]. This is mainly due to advances in the definition of chromosomal markers, and in related sequencing technologies. Advanced study-designs and statistical methods have shifted the focus of genetic research from studying rare disorders within families, to studying common and complex disorders segregating within large populations. The new methods have allowed us to dissect the genetic architecture of complex disorders including the identification of the causal genomic loci, estimation of the disease heritability, estimation of effect sizes of different loci and their non-additive interactions.

Linkage analysis

The earlier breakthroughs in linking genotype with phenotype involved studies of Mendelian disorders that can be mapped to a single gene and a single mutation. These studies were often family-based and used linkage analysis to define the candidate chromosomal region, followed by positional cloning to narrow down the candidate region to a causal gene. Genetic variation within families was used to construct linkage maps, and risk loci and genes were mapped to a particular chromosomal location by testing for co-segregation with small panels of markers—often just a few hundred. The first successful application of this approach identified genomic loci responsible for an X-linked phagocytic disorder—chronic granulomatous disease (*CYBB*) [20]. This was soon followed by identification of the loci and ultimately genes responsible for other genetic disorders including Duchene muscular dystrophy (*DMD*) [21], cystic fibrosis (*CFTR*) [22], Huntington disease (*HTT*) [23,24], polycystic kidney disease (*PKD1, PKD2* and *PKHD1*) [25-27] , phenylketonuria *(PAH)*[28], albinism (*TYR*) [29] and many more. Currently, Online Mendelian Inheritance in Man (OMIM) catalogues 4,500 human disorders for which the underlying genetic mutations are known (http://omim.org/statistics/entry).

The success of linkage analysis and positional cloning was mainly limited to the identification of high-penetrance monogenic variants that mainly disrupt the structure of proteins such as huntingtin [30]. However, most genetic disorders including cardiovascular diseases, diabetes and neurodegenerative diseases are actually the result of combination of inherited variants in multiple genes that have small or moderate effects and that often do not modify protein structure at all [31]. For example, a recent study has identified 108 loci that are significantly associated with schizophrenia none of which produce known protein differences [32]. For these types of complex polygenic diseases, family-based linkage studies often suffer from unavailability of family of multiple generations or sufficient numbers of genetically informative families, particularly for late-onset diseases such as Alzheimer’s and Parkinson’s. Linkage analysis also suffers from poor genetic resolution (typically on the order of a few centimorgans). Additionally, for polygenic disorders it is highly unlikely that every family will be segregating for the same collection of causal variants (genes) and combining data may adversely affect the linkage analysis.

Genome wide association studies

The development of the common disease/common variants hypothesis in mid-1990s led to the idea of genome-wide association studies (GWAS) [33,34]. This hypothesis predicts that common disease-causing alleles will be found in all human populations which manifest a given disease. In other words, common complex diseases could be accounted for by a few common variants (minor allele frequency > 5%) with moderate effects. This led to large-scale human genotyping initiatives such as the International HapMap Project [35,36] to catalog common genetic variants that are segregating within and between human populations. The Phase III of this project genotyped 1.6 million common single nucleotide polymorphisms segregating in 11 human populations, representing multiple ethnicities. Another large-scale genomics initiative, 1000 Genomes Project, has been using a combination of whole-genome sequencing, deep exome sequencing, and dense microarray genotyping to establish the most detailed catalogue of genetic variants (minor allele frequency > 1%) in human populations [2,5]. The third phase of 1000 Genomes Project has just finished and generated a haplotype map of 84.7 million SNPs, 3.6 million short insertions and deletions (indels), and more than 60,000 larger deletions segregating in 26 human populations [37].

A typical GWAS examines a large number of common genetic variants (500,000 to five million SNPs) in a large number of case and control individuals to identify alleles associated with a trait or disease. Typically, thousands of individuals are genotyped in both case and control groups. Allele frequencies for genotypes (SNPs) between two groups are compared to reveal genotypes that are overrepresented in cases compared to controls and therefore likely to be associated with disease risk variants. The first successful GWAS study investigated age-related macular degeneration (AMD) to identify variants in complement factor H (*CFH*) as major risk factors [6]. Since then, there has been a deluge of GWA studies (GWASs) that have identified thousands of statistically significant SNPs associated with common diseases including coronary artery disease [38], type 1 and 2 diabetes [8,39,40], bipolar disorder [41], hypertension [7,42] and many more. Currently, the GWAS catalogue [9] at NHGRI-EBI (http://www.ebi.ac.uk/gwas/) consists of over 2,000 GWAS publications linking nearly 4,500 statistically significant loci ($p < 10^{-8}$) to over 500 human traits and diseases. Extensive cataloguing of common variants and high-throughput genotyping has made GWAS a well-established method to identify genetic variants for the complex diseases.

Unlike family-based studies GWAS offers the advantage of exploiting unrelated individuals, making the task of data collection much easier. However, population-based studies are confounded by population stratification—presence of systematic difference in allele frequencies between subpopulations due to their different ancestry. As a result, any SNP with a considerable variation in its frequency between different subpopulations could be spuriously associated with the disease of interest if these subpopulations themselves considerably differ in the prevalence of the disease [43]. A number of statistical approaches have been developed to capture and control for the complex population structures in GWAS [44-48].

Electronic medical records

The last decade has also seen rapid growth in efforts to implement electronic medical records (EMRs) systems in large health care systems. EMR data sets contain rich and diverse medical information including patient history, diagnoses, prescribed medications, and quantitative data from many possible clinical tests. The structured and unstructured information (free-text clinical notes) in EMRs can be extracted using the defined International Classification of Diseases, version 9 and 10-CM codes [49-51] and natural language processing approaches. Longitudinal EMR datasets have been used for observational-based healthcare research to improve patient care [52,53].

Phenome-wide association study (PheWAS): advantages and challenges

Lately, EMR-derived phenotypes have been linked with genotypes to perform phenome-wide association study (PheWAS) [54-59]. Large scale biorepositories such as the electronic Medical Records and GEnomics (eMERGE) network [60-65] have been systematically integrating massive EMR datasets (phenotypes) and DNA biorepositories (genotypes) for high throughput genome-to-phenome research.

Unlike GWAS that uses a phenotype-to-genotype approach, PheWAS uses a reverse genetic approach to perform genotype-to-phenotype association. It starts with a known genetic variant and tests for its association over a wide spectrum of clinical phenotypes derived from EMRs. The first proof-of-principle PheWAS study was published in 2010 by Denny *et. al* [59]. They selected five known disease-associated SNPs (previously identified by GWAS) and genotyped a cohort of 6,005 patients for these SNPs. Each of these SNPs was then associated with hundreds of disease codes that comprised the phenome. The PheWAS recaptured four of the five known SNPs-disease associations and also found novel associations with other diseases. For example, a SNP (*rs3135388*) that was previously associated with multiple sclerosis was also found to be associated with erythematous conditions. Several recent studies have successfully exploited PheWAS to identify novel SNP-disease associations [54,57,59,64,66,67]. Whereas GWAS focuses on only one disease at a time, PheWAS has the ability to measure genetic associations with potentially thousands of traits and diseases simultaneously, enabling systematic detection of genetic variants with pleiotropic effects [56,58,67]. This can aid in the repurposing of approved drugs and save enormous amount of time, money and efforts that go into the therapeutic development and testing process. For example, variants in *TNFSF11*, a tumor necrosis factor, are associated with Crohn's disease. Denosumab, a monoclonal antibody that targets *TNFSF11* is currently marketed for the treatment of postmenopausal osteoporosis, but it may also be considered for Crohn's disease [68].

A prerequisite of conducting a PheWAS is the availability of massive phenome data. As a result, all PheWAS studies have exploited disease phenotypes derived from EMR datasets. However, in practice PheWAS can be applied to any well-characterized cohort that have been extensively genotyped and for which deep phenomes have been assembled, such as the Framingham Heart Study cohort or the Avon Longitudinal Study of Parents and Children cohort [69-74]. As I show below it is also possible to conduct a PheWAS using deeply phenotyped cohorts of model organisms.

GENETIC ASSOCIATION STUDIES IN MOUSE MODEL SYSTEMS

The laboratory mouse has been extensively used as a model organism to dissect the genetic architecture of complex phenotypes and disorders. Studying complex phenotypes in mouse offers several advantages over human. First, it is practical to acquire cellular and molecular traits from disease-relevant tissues and cell types. Second, it is possible to replicate experiments in reference cohorts (also known as reference panels or reference populations), which is impossible in humans except for in cases of monozygotic twins. Third, it is easy to control the environment and model gene–environment (GXE) interactions in mice [75]. Fourth, despite strong functional effects, the minor allele frequencies are often too low in the human population to attain sufficient statistical power and significance in large association studies. In contrast, most of murine crosses have been derived from two inbred strains, and as a result allele frequencies are close to 0.5. Fifth, it is possible to perform knockins, knockouts, and knockdowns in mice to identify the causal gene from the list of candidate genes. Lastly, genetic research in humans has added complexities including confidentiality constraints and higher ethical limitations on experimental protocols. Genomic information derived from human participants especially those in diseased groups is sensitive, and strict guidelines and policies (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html) must be followed to protect the privacy and confidentiality of participants and their descendants [76]. However, one of the main disadvantages of murine cohorts is that linkage disequilibrium is typically at least an order of magnitude larger than in humans. As a result, mapping studies using currently available murine populations often identify large genomic regions with tens to hundreds of candidate genes.

Mapping populations in mouse model systems

Mice cohorts, including F2 intercrosses and backcrosses, heterogeneous stocks [77], sets of recombinant inbred (RI) strains (e.g., BXD [78], LXS [79], and the Collaborative Cross [80]) have been extensively used for mapping of both Mendelian and complex traits. These crosses differ greatly in their genetic diversity, mapping power, and resolution [81,82]. RI strains are generated by intercrossing two or more parental inbred strains, followed by repeated sibling matings for at least 20 generations. The repeated mating of siblings for 20 generations or more generates fully inbred strains. Each RI strain represents a unique and fixed chromosomal mosaic of the parental genomes. Once all members of a large set of RI strains have been fully inbred and genotyped, then the set can be used as an immortal and genetically defined resource—a genetic reference panel. The historical disadvantage of RI strains was their limited numbers and modest power and precision of associated QTL studies. Throughout most of the 1990s there were fewer than 30 strains per family. Now however, three mouse RI GRPs consist of 60 to 150 strains each (n ~150 for BXD, n ~ 60 for LXS, and n ~100 for the Collaborative Cross) [79,83]. The main disadvantage of RI strains is not QTL mapping power or precision, but steadily rising costs of acquisition and maintenance of large numbers of strains.

BXDs—a replicable experimental cohort for genomic medicine studies

The BXD family is made up of ~150 RI strains (some still in progress) that descend from intercrosses between C57BL/6J and DBA/2J strains (**Figure 1-1**). Currently, they are the largest and oldest RI set available [84]. In chapter 2, I have compared the genome sequences of the parental strains of the BXDs to identify ~5 million sequence variants that are segregating in this family.

The BXD genetic reference panel has been used to study higher-order complex traits in diverse research domains including neurobiology, physiology, pharmacology and immunology since the mid-1970s and genetics of gene expression since the early 2000s. As a result, they have the largest coherent multiscalar phenotype data set (aka "phenome") for any segregating population, consisting of 5000 diverse phenotypes (www.genenetwork.org) and many gene expression data sets. Because these strains are genetically immortal they can be used as a replicable experimental cohort for personalized genomic medicine studies. The genetic immortality enables assembly of deep phenomes for the same set of individuals (strains) over time. Additionally, matched cohorts can be raised in under different environments to study gene-by-environment interactions. Approximately 120 BXD progeny lines have now been extraordinarily well genotyped (30 more in progress) and can be exploited to relate sequence variants to phenotypic differences.

Over the last three decades, the BXD family has been exploited mainly using the forward genetic approach such as QTL mapping. This approach starts with heritable differences in phenotypes and defines loci and causal variant. A few recent achievements using this approach include identifying *Ubp1* for blood pressure [85], *Aplp* for hypophosphatasia [86], and *Mrps5* for longevity [87]. In contrast, a reverse genetic approach such as PheWAS starts from known sequence variants and identifies downstream phenotypic effects. A few recent achievements include *Comt* for a number of neuropharmocological traits [88] and *Per3* for stress/anxiety traits [89]. The availability of a comprehensive catalogue of genetic variants linked to the deep phenome datasets make BXDs highly practical to perform the first PheWAS (genome-wide reverse genetic scan) using a model organism.

GENE EXPRESSION AS MOLECULAR PHENOTYPES

Sequencing of large cohorts has now become fairly straightforward. However, generating variant data for thousands of individuals has limited predictive value unless integrated with a wide-spectrum of phenotype data. Large-scale phenotyping especially of the clinical traits is still intractable due to high cost, difficult implementation, and

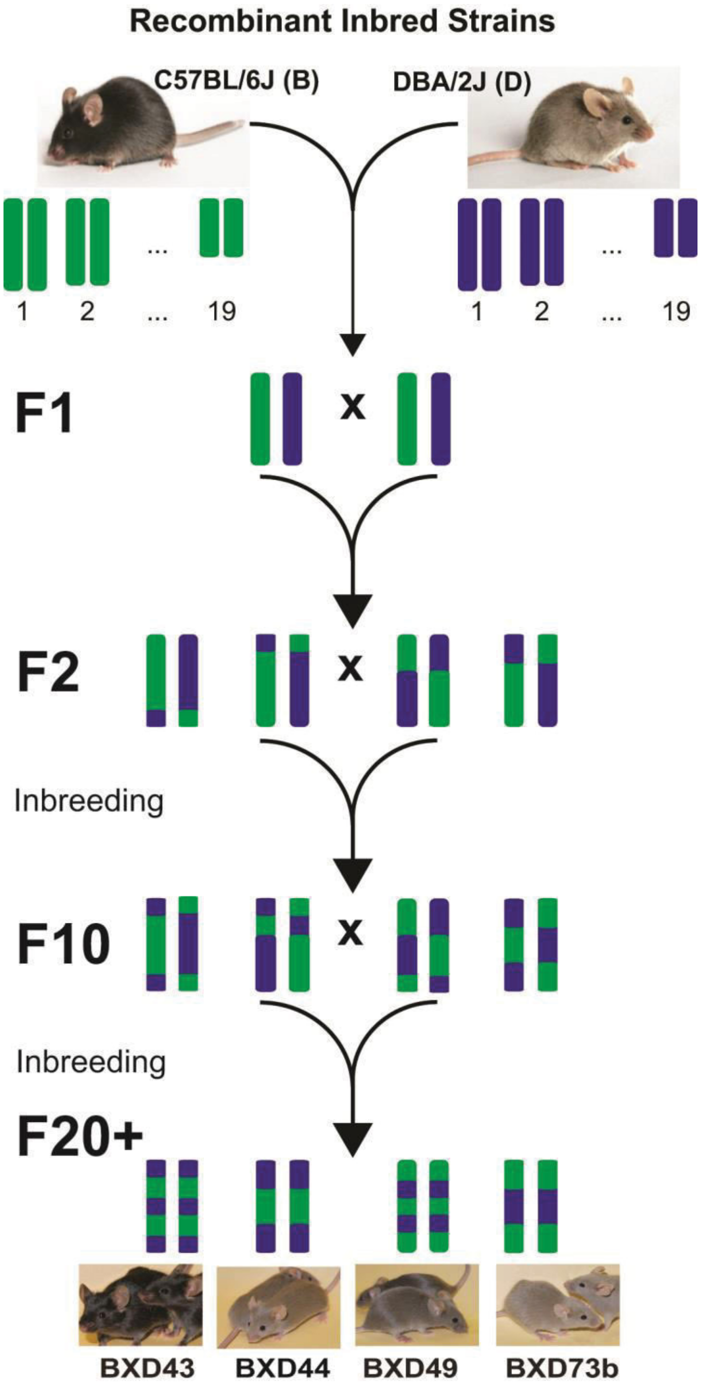


Figure 1-1. Derivation of the BXD family

Notes. The parental strains are crossed to generate F1 progenies consisting of genetically identical individuals. F1 individuals are intercrossed to generate F2 individuals. In the F2 population, each individual has a unique genotype due to the recombination of the alleles from the heterozygous F1 parents. Repeated sibling mating is performed for 20 generations to generate inbred strains.

ethical restrictions. An alternative is to use molecular phenotypes including transcript, protein and metabolite abundances and their modifications to study the genetic basis of differences in disease risk. Gene expression levels are highly replicable and reproducible and have been widely used as reliable prognostic indicators of diseases [90-92]. Genetic linkage and association studies on gene expression levels have demonstrated high heritability [93] indicating that genetic variants often confer disease risks by affecting gene expression.

The analyses of complex phenotypes in the pre-genomic era focused on coding variants—including nonsense, missense, and frameshifts variants,—but GWAS studies conducted over the last decade have demonstrated that a large majority (>90%) of trait/disease-associated variants are located in non-coding regions of the human genome [9]. These non-coding variants act by modulating gene expression, and they are the major causes of variation in susceptibility to complex diseases [94-96]. Similar to classical phenotypes, gene expression can exhibit Mendelian or multigenic inheritance patterns and are therefore amenable to association studies. The control of expression is usually genetically complex (polygenic) and large numbers of other genes and sequence variants can potentially influence expression of the target transcript or protein. For example a group of cooperating transcription factors may control expression of a key transmitter receptor or an ion channel. These effects give rise to so-called *trans* eQTLs that map far from the target gene itself—usually on different chromosomes (**Figure 1-2a**). In contrast, expression of mRNAs may also be controlled by sequence variants that are in or very near to the parent gene itself (**Figure 1-2b**). For example, a polymorphism in a promoter, enhancer, splice acceptor site, or the 3' UTR of a gene may produce differences in transcriptional rates, mRNA stability, or ratios of alternative transcripts. When mapping the expression of mRNAs or proteins, this type of genetic "self-control" produces so-called cis-acting QTLs or *ci*s eQTLs [97]. In short, *cis* eQTLs are first-order local effects, whereas *trans* eQTLs are second-order distant effects.

CURRENT STATUS OF GENOMIC MEDICINE

Current successes of genomic medicine include clinical diagnosis of monogenic diseases and disorders, improved therapeutic efficacy and safety of drugs, drug repurposing, and molecular characterization of cancers to select more effective treatments. However, with the exception of above mentioned clinical applications, genomic medicine has yet to be embraced to the extent that was initially anticipated following completion of Human Genome Project [98]. Here we discuss a few examples of how the early phase of genomic medicine is impacting healthcare.

Better diagnoses and early interventions

Diagnostic kits that allow screening of genetic carriers for disorders including breast and ovarian cancer (*BRCA1* and *BRCA2*) [99-102], colon cancer (*MLH1*, *MSH2*, *MSH6* and *PMS*2) [103-106], melanoma (*CDKN2A*) [107], rheumatoid arthritis

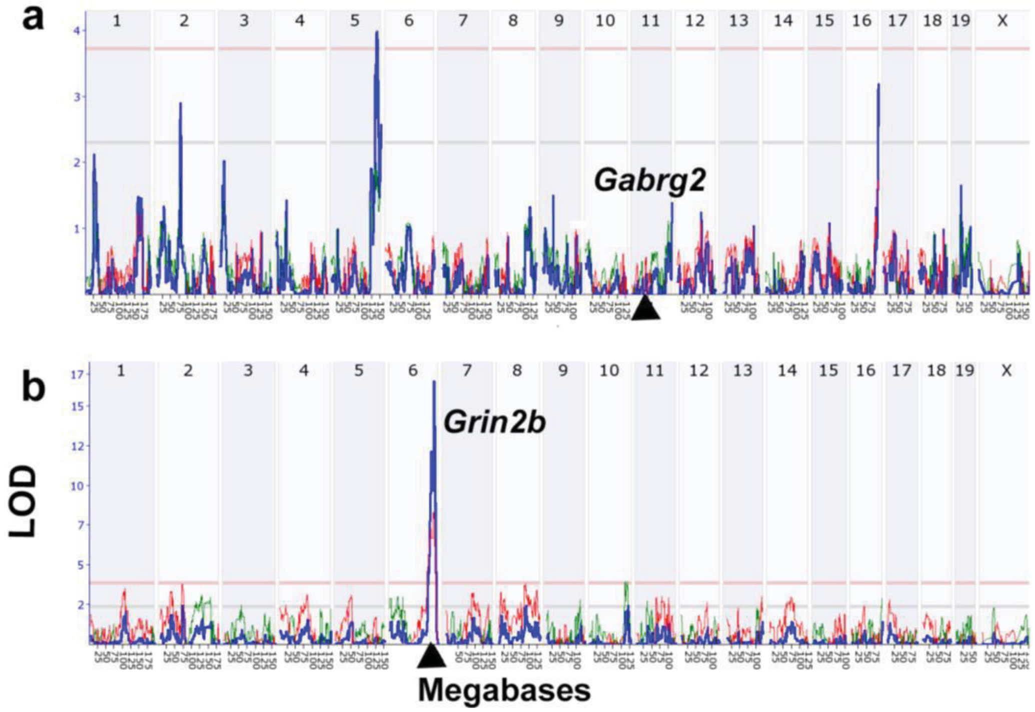


Figure 1-2. Linkage maps of cis and trans eQTLs in mouse hippocampus

Notes. **(a)** *Gabrg2* expression is controlled by a trans eQTL on Chr 5 at 138 Mb (LOD = 3.94 on the Y axis). The *Gabrg2* gene itself is located on Chr 11 at 41 Mb (triangle on X axis). **(b)** In contrast, *Grin2b* expression is controlled by a cis eQTL with a peak LOD score of 16.73 located on Chr 6 at 135 Mb. This location corresponds precisely to the location of the *Grin2b* gene (triangle). The horizontal lines provide genome-wide significance thresholds for the QTL determined by permutation analysis (upper <.05 and lower <.63). All data here were generated in GeneNetwork (www.genentwork.org) using the BXD mouse *Hippocampus Consortium M430v2 (Jun06) PDNN* array data set (GeneNetwork.org, accession number GN112, *n* = 67, probe sets 1418177_at and 1457003_at).

(*HLA-DR4*) [108], cystic fibrosis (*CFTR*) [22], and thrombophilia (*FV, FII, MTHFR*) [109-111] have been widely used to guide preventive care. For example, prophylactic mastectomy or oophorectomy is recommended to predisposed individuals and has shown to reduce the risk of cancer by 90-95% in women [112]. Similarly, genetic screenings are available for prenatal and newborns to detect birth defects and genetic diseases including cystic fibrosis, severe combined immunodeficiencies, phenylketonuria, tyrosinemia, sickle cell anemia, hearing loss, and congenital heart defects. Currently, testing of 32 core disorders and 26 secondary disorders is recommended by the U.S. Department of Health and Human Services Secretary's Advisory Committee on Heritable Disorders in Newborns and Children [113].

Drug response and dosage

Diagnostic kits based on pharamacogenomic markers that type the cytochrome P450 family—a major subset of all drug metabolizing enzymes in liver— have been designed to determine therapeutic strategies and effective dosage. For example, *CYP2D*, a member of the cytochrome P450 gene family, is responsible for the metabolism of approximately 25% of all clinically used drugs including codeine, oxycodone and tramadol (pain), tamoxifen (breast cancer), dextrometorphan and quinidine (neurological disorders), tetrabenazine (Huntington's disease), atomoxetine (Attention-deficit/hyperactivity disorder), citalopram and desipramine (depression), and fluvoxamine (obsessive compulsive disorders) [114]. Copy number variants (CNVs) in *CYP2D6* control variation in drug response among individuals [115]. Individuals with multiple copies of *CYP2D6* are ultra-rapid metabolizers. They quickly convert codeine into morphine and may experience potentially dangerous opioid effects. In contrast, poor or slow metabolizers suffer from poor analgesia.

Accurate classification and customized treatment plans for cancers

High throughput sequencing can accurately determine driver mutations or genes associated with heterogeneous cancers and aid in a better understanding of disease pathology, as well as customization of treatment plans. Screening mutations in non-small cell lung cancer related genes including *EGFR*, *ALK*, *HER2*, *KRAS*, *BRAF,* and *PI3KCA* helps select targeted therapies based on the driver mutations [116]. For example, Dabrafenib, an inhibitor of the BRAF protein has shown to be effective against lung cancer associated with mutated *BRAF* gene [117]. Targeted treatments based on driver mutations have shown to increase survival rates compared to non-targeted treatments [118,119].

BIOINFORMATIC CHALLENGES FOR GENOMIC MEDICINE

Processing and managing of high-throughput sequence data

High throughput sequencing offers several advantages relative to array-based genotyping or expression assays. First, unlike genotyping arrays, whole genome sequencing is not limited to interrogating only known sequence variants. Similarly, RNA-sequencing (RNA-seq) enables expression quantification of novel transcripts that are not represented on arrays. Second, whole genome sequencing makes it possible to detect large and complex structural variants. These variants are not as common as SNPs but have significant effects on expression. However, there is a high bioinformatics overhead required to store and process sequencing data. Sequencing a mammalian genome at 100x coverage can easily generate up to half a terabyte of raw sequence data. Moreover, size of the intermediate data generated during the analysis can get nearly double the original size. Thus, large-scale storage infrastructure and high network bandwidth connection are essential for massive sequencing projects. Better compression methods to minimize storage costs are an area of active research. New compression methods including CRAM, Goby and HDF5 [120-124] to store genomic alignments have been developed. However, they have not been fully embraced by the next-generation sequencing community due to their incompatibility with most of the current analysis tools. Unlike arrays, the computational workflows for high throughput sequencing data analysis are too intensive to be performed on a desktop computer and require high performance computing clusters with hundreds of processors. Adaption of cloud computing strategies has been on the rise and enables users to customize hardware and computational power based on the project requirements.

Interpretation of the functional impact of the genomic variants

Functional interpretation of genetic variants is crucial to prioritize candidate variants in association studies. Functional interpretation of coding variants is relatively straightforward and their impact can be assessed by annotating them against known gene models. However, our ability to interpret the impact of non-coding regulatory variants is highly limited despite their known roles in various diseases [95,96]. Early studies to evaluate the impact of non-coding variants mainly exploited sequence conservation across multiple species to quantify likely evolutionary constraints on the variant position [125-130]. Genome wide scans using position weight matrices [131] have also been used to identify if variants overlap any known transcription factor binding motifs. However, Schmidt and colleagues have observed large interspecies differences in transcriptional factor binding regions [132]. Additionally, short matrices have low sequence specificity and may identify numerous false positive binding sites across the genome [133]. Recent studies have adopted multi– and integrated –omics approaches that incorporate a wide range of annotations to evaluate functional impact of non-coding variants. Khurana and colleagues [134] used allele frequencies of sequence variants from 1000 Genomes Project [2] to distinguish deleterious variants from the neutral variants. They proposed that the neutral variants have higher derived allele frequency at the population level and deleterious variants get removed by purifying selection. They then overlapped this information with functionally non-coding elements identified by Encyclopedia of DNA Elements (ENCODE) project to identify functionally important non-coding variants. Combined Annotation Dependent Depletion (CADD) and Genome Wide Annotation of Variants (GWAVA) are machine learning based approaches that work along the same line by integrating diverse resources to identify functional non-coding variants [135,136]. Recently, large numbers of genome-wide studies have successfully integrated high-throughput omic measurements, including gene expression and epigenetic variation data to increase the power for discovery of causal genes and to better understand the possible molecular and cellular mechanisms of the disease [137-140].

ORGANISATION OF THE DISSERTATION

The work presented in this dissertation is a genome-wide reverse genetics analysis. We have exploited a large family of recombinant inbred mouse strains—the BXD cohort to perform a phenome-wide association study to investigate genome-to-phenome relations at multiple scales—from mRNA and protein levels to disease risk, behavior, and environmental interactions.

The second chapter explores the genomic variation between parental strains of BXDs using deep (~100X) sequence data of the DBA/2J inbred strain. This analysis revealed considerable genomic variation including ~4.46 million SNPs, ~0.94 million indels, and 20k structural variants segregating in the BXD family.

The third chapter examines the functional impact of genomic variation on gene expression using transcriptomic data from isogenic hybrids (C57BL/6J X DBA/2J F1s). This analysis revealed that *cis* acting variation in expression is pervasive and is detected in roughly 50% of all assayable genes in liver. Genes exhibiting high allelic differences in expression in conjugation with high-impact coding variants should be key molecular resources for reverse genetics analysis.

The fourth chapter examines the functional impact of genomic variation on a wide spectrum of high-order phenotypes [141,142] and molecular phenotypes across multiple tissues [86,142,143]. We successfully replicated almost all of the known genome-to-phenome associations in BXDs, and also identified a few novel associations. We exploited a large human clinical cohort—the Vanderbilt BioVU cohort— for validation and cross species translation of the novel associations. We demonstrate that phenome scans can be effective at linking sequence variants to a range of phenotypes and can be used to identify novel genome-to-phenome relations or validate hypothesized associations from independent studies.

Finally, the fifth chapter summarizes and discusses the main results of my dissertation.

CHAPTER 2. SEQUENCING AND CHARACTERIZATION OF THE DBA/2J MOUSE GENOME

SYNOPSIS

The DBA/2J mouse is one of the oldest and widely used inbred strains. It exhibits many unique anatomical, physiological, immunological and behavioral phenotypes. In addition, it is one parent of the large BXD family of recombinant inbred (RI) strains–a widely used murine genetic reference population. The genome of the other parent of this BXD family—C57BL/6J—has been sequenced and serves as the mouse reference genome. We sequenced and analyzed the genome of DBA/2J to generate a comprehensive catalogue of ~5.4 million sequence variants, relative to the reference genome. These variants segregate in the BXD family, presently comprising of 120+ RI strains. The variant data can be exploited to initiate reverse genetic analysis of complex traits, particularly by exploiting high-impact variants including nonsense, frame-shift, splice-site, radical missense, copy number, and large insertion and deletion that differentially affect members of the BXD family. The variant catalogue is also essential for unbiased alignment of RNA-seq and ChIP-seq data generated using BXD strains and any other cross involving DBA/2J as a parental strain.

INTRODUCTION

The DBA inbred strain has the distinction of being the oldest of all inbred strains. It was first developed by C.C. Little in 1909 by inbreeding from a stock of mice segregating for coat color. During 1929-1930, DBA substrains were crossed to establish new substrains including DBA/1 and DBA/2. The DBA/2 strain was transferred to G. B. Mider in 1938, and subsequently transferred to the animal facilities of Jackson Laboratory (J) and National Institute of Health (N) in 1948 and 1951 respectively. Since then, they have been maintained separately as DBA/2J and DBA/2N lines (**Figure 2-1**).

DBA/2J is one of the most widely used inbred strains and exhibits many unique anatomical, physiological, immunological and behavioral phenotypes. A number of these phenotypes closely mimic human diseases and disorders. A few well known age-related phenotypes are progressive eye abnormalities and hearing loss are caused by mutations in *Gpnmb* [144], and *Tyrp1* [145,146] and *Cdh23* [147,148] genes respectively. The unique characteristics of DBA/2J are often contrasted with those of the C57BL/6J inbred strain that serves as the mouse reference genome. The two strains are genetically highly divergent and show a wide variety of phenotypic differences. For example, C57BL/6J and DBA/2J have high and low susceptibility to diet-induced atherosclerosis [149-151]; high and low preference for alcohol and morphine [152-158]; high and low resistance to influenza infection [159-161]; low and high bone mineral density [162-165]; low and high susceptibility to audiogenic seizures [166-169]. The high levels of genetic and phenotypic variation between these strains have made them highly favorable to be used for genetic linkage studies. Starting in early 1970s, Benjamin A. Taylor at the Jackson

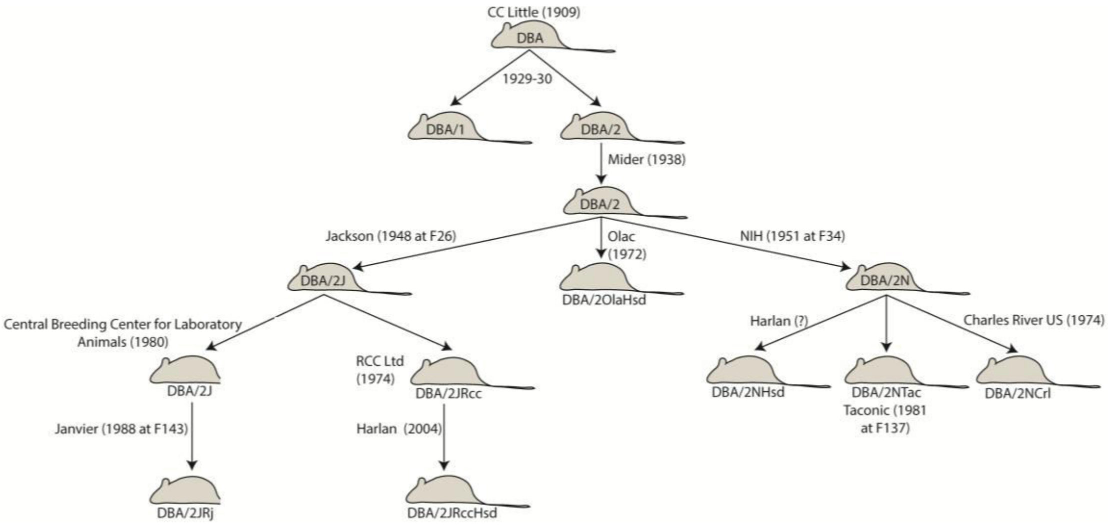


Figure 2-1. DBA/2 breeding history

Notes. DBA/2 is the oldest inbred strain and originated as part of breeding efforts by C. C. Little around 1930. The DBA/2 strain was transferred to G. Burroughs Mider in 1938, and subsequently transferred to the animal facilities of Jackson Laboratory (J) and National Institute of Health (N) in 1948 and 1951 respectively. Since then, numerous DBA/2 substrains were created by separation and breeding by different vendors, leading to genetic drift. These nearly identical lines create a valuable genetic resource for studying the downstream effects of spontaneous and naturally occurring mutations. Information regarding substrain derivation dates was compiled from individual vendor Web sites.

laboratory started intercrossing C57BL/6J and DBA/2J to produce the BXD family of recombinant inbred strains [170]. Since the early-1970s, the BXDs have been used extensively to study the genetic basis of complex traits [84]. They have also been used to study the genetic basis of gene expression since the early 2000s after the advent of high-throughput expression arrays. BXDs are currently the largest (120+ lines and expanding to ~150) and the best phenotyped genetic reference population. More than 5,000 phenotypes for these strains have been measured and published, and all of these phenotypes are accessible at GeneNetwork (www.genenetwork.org).

Whole genome sequencing now makes it practical to exploit reverse genetics to simultaneously study functions of thousands of sequence variants that segregate in intercross and backcross progenies generated by mating completely sequenced parental lines [171] —the BXDs being a prime example. However, the analysis of high-throughput short read data to discover sequence variants is computationally challenging, and susceptible to errors associated with mapping artifacts and sequencing chemistry. We followed the best practices for variant discovery to provide the accurate and most detailed picture of molecular variation between these two genomes to date. In this chapter, we describe a comprehensive and high confidence catalogue of SNPs, indels, and structural variants for the DBA/2J mouse strain, the father of the BXD family, relative to C57BL/6J, the reference strain for all mouse genomics and the mother of the BXD family. We annotated the sequence and structural variants against mouse gene models to assign functional consequences. The detailed catalogue of genetic variation segregating in the BXD family is essential to 1) link high-impact variants with high-order phenotypic differences; 2) correct for allelic bias in RNA-seq and ChIP-seq read mapping; 3) assay differential expression of alleles in isogenic heterozygous F1 individuals.

MATERIALS AND METHODS

Genomic data for DBA/2J

We downloaded sequencing data for DBA/2J from the European Nucleotide Archive, accession numbers ERP000044 and ERP000927 [172]. It consists of eleven paired-end libraries sequenced on the Illumina GAII. Average read lengths and insert-sizes varied between 54–100 nt and 150–600 nt respectively. We also downloaded sequencing data from the Sequence Read Archive, accession number SRP001135 [173]. This data was generated at UCLA as a part of the DBA/2J sequencing study at UTHSC. It consists of three paired-end libraries sequenced on the Illumina GAII with read length of 100 nt and insert-sizes between 250–350 nt.

Read alignment and post-alignment processing

We organized and processed the sequencing data at multiple levels including raw data for each sequencing lane, aligned reads for each library and combined aligned reads for the DBA/2J samples. A variety of post-processing steps were performed on the data in a sequential manner to improve the alignment and accuracy of variant calls.

Lane level. Reads were trimmed to remove low quality base calls ($q < 20$) using Trimmomatic [174] and trimmed reads shorter than 40 nucleotides were removed. Reads were aligned to the C57BL/6J reference genome (mm10) using Burrows Wheeler Aligner (BWA version 6.1) [175] with default parameters to generate sequence alignment/map (SAM) files [176]. SAM files were converted to binary SAM (BAM) format and their headers were appropriately modified to include generic information such as lane identifier, library identifier etc. using Picard 'AddOrReplaceReadGroups'(version 1.67, http://picard.sourceforge.net). Base qualities of aligned reads were recalibrated using Genome Analysis Toolkit (GATK, v2.1-9) 'TableRecalibration' [177]. This tool empirically models errors in assignment of base qualities by the sequencing machine and generates new qualities that are highly accurate.

Library level. All lanes (BAMs) for each library were merged into one BAM file using Picard 'MergeSamFiles'. PCR duplicate reads (those that map to the same genomic location as a pair) except the one with the highest sum of base qualities were flagged using Picard 'MarkDuplicates'. Duplicate reads are the result of library amplification and should not be considered as independent evidence supporting a variant call. Additionally, PCR-induced errors can easily propagate to subsequent duplicate reads causing spurious SNPs.

Sample level. BAM files for each library were combined together to create a master file containing all DBA/2J sequences. Finally, GATK 'IndelRealigner' was used to perform local realignment of reads around indels from the Mouse Genome Project [172] as well as putative indels identified by GATK. Reads aligning on the edges of indels in the reference genome often aligned with mismatches due to the fact that the penalty for opening a gap is higher than that of incorporating a mismatch. Local realignment of reads around indel removes spurious SNP calls due to this alignment artifact.

Variant calling pipeline

SNPs and indels. GATK 'UnifiedGenotyper' and Samtools (version 0.1.18) 'mpileup' function combined with BCFtools were used to discover SNPs and small indels in the variant call format (VCF). Reads with bitwise flag (0x704) that include reads flagged as 1) duplicates (0x0400), 2) failed QC (0x0200), 3) non-primary alignment (0x0100), and 4) unmapped (0x0004) were not considered for variant calling. GATK 'UnifiedGenotyper' was used with the parameters '-bfh 4000 -rbs 1000000 -dcov 500 --heterozygosity .0001 --genotype_likelihoods_model BOTH'. Samtools 'mpileup' was used with the parameters '-C 50 -d 1000 -E -Q 20 -q 20', followed by BCFtools with default parameters. We only selected variants that were jointly identified by GATK and SAMtools

Structural variants. Structural variants were identified using three different approaches including split-read, discordant read pairs, and uneven distribution in the read depth.

Split-read. Pindel (version 0.2.4) [178] was used with the parameters '-v 500' and 'pindel2vcf' was used to generate vcf files. Pindel exploits read-pairs where one read is uniquely aligned and its mate read could only be aligned partially. The split-alignment of the read indicates that it either overlaps deletion breakpoints or contains one end of an insertion.

Discordant read pairs. BreakdancerMax (version 1.1) [179,180] was run with the parameters '-c 3 -m 10000000 -q 25 -r 3 -h –f' to detect insertions, deletions and inversions. It uses anomalous read pairs that either do not align in concordance with the library design or align considerably closer or farther than the expected insert sizes.

Uneven distribution of read depth. Copy Number Detector (CND, version 1.3) [181] was run with the default settings. CND is exclusively designed to detect copy number gains and losses in homozygous diploid organisms, such as inbred mouse strains. It uses both the coverage of sequence reads, and the rate of apparent heterozygous SNPs (paralogous sequence variants) to determine CNV gains and losses.

Variant filtering pipeline

In-house Python scripts were used to filter raw variant calls based on multiple criteria. These filters removed low-quality variants due to alignment artifacts and sequencing errors. They include number of reads supporting the variant, alignment quality of reads supporting the variant, quality scores of variant calls by sequencer etc. DBA/2J is a fully inbred strain and we therefore only retained homozygous SNPs and indels. Structural variants that were supported by less than three libraries were removed. Variant quality thresholds of 30 and 50 were used to filter Pindel and Breakdancer variant calls respectively. We removed structural variants that overlapped low complexity regions including centromere, telomere and assembly gaps in the reference genome. The genomic coordinates of these regions were obtained from the UCSC Table browser [182] (mm10). Python scripts along with the default filtering parameters that were used to filter raw variants calls from various variant detection tools including GATK, Samtools, Pindel and BreakDancerMax can be downloaded here:
https://github.com/ashutoshkpandey/Variants_call

Variant annotation pipeline

We used SnpEff [183] to perform functional annotation of SNPs and indels, and categorize them into nonsense, splice-site, frameshift and missense SNPs. The annotation was performed against mouse gene models from Ensemble (version 69) and RefSeq (downloaded in October 2012). An in-house Python script (https://github.com/ashutoshkpandey/Annotation/blob/master/Grantham_score_calculator.py) was used to generate Grantham scores to predict the impact of the amino acid substitution due to missense mutation. Missense variants categorized as 'radical' and 'moderately radical' were defined as deleterious. We also used PolyPhen-2 (version 2.2.2) [184] with default HumDiv model to predict the impact of the amino acid substitution. Mouse annotation and sequence database for Polyphen-2 were prepared by downloading mouse Uniprot (version 2012_11) [185] and PFAM (downloaded in November 2012) [186]. Missense variants predicted to be 'probably damaging' or 'possibly damaging' were defined as deleterious.

Experimental validation for variants

SNPs and indels were selected for validation by traditional Sanger sequencing. Primers were designed using Primer3 (http://frodo.wi.mit.ed/primer3/). PCR assays were performed using 5 ng DBA/2J genomic DNA, 10 pmol each of forward and reverse primer in 50 µl. The following cycle parameters were used: 95 °C for 4 min, 35 cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min, and 72 °C for 5 min. PCR products were purified with 2 µL ExoSAP–IT (Invitrogen Corporation). Sanger sequencing was performed using an ABI 3730.

RESULTS

Whole genome sequencing and alignment

The DBA/2J mouse genome was sequenced to a depth of 90-fold coverage by the Illumina HiSeq 2000 sequencing platform. 14 paired-end libraries with different insert-sizes ranging between 150-600 nt were sequenced to generate 4.08 billion reads (374 billion nucleotides). The sequencing reads were aligned against the GRCm38 reference genome (mm10), followed by base recalibration, flagging of duplicates reads and local realignment of reads around indels (**Figure 2-2**). A total of 3.56 billion reads were aligned onto the reference genome, and 92% of the aligned reads were properly paired. We defined 'accessible reference genome' as fraction of the reference genome excluding gaps and regions of low complexity (ambiguous read alignment). Around 90% and 80% of the 'accessible genome' has coverage of 50 and 70 reads per base respectively. Over 22 percent of the total aligned reads were not considered towards coverage analysis because they belonged to one of the following categories: PCR duplicates, reads with low base quality, and reads with low mapping quality.

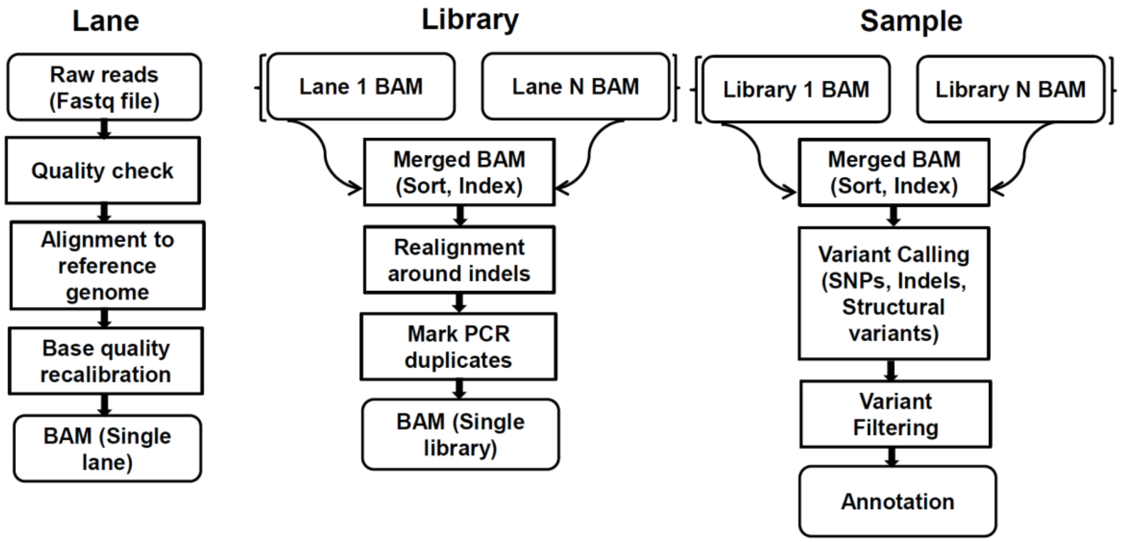


Figure 2-2. Variant discovery workflow

Notes. Sequence data is organized and processed at three levels including raw read data for each sequencing lane, aligned reads for each library, and combined reads for each sample. The first step involves quality control assessment, alignment to the GRCm38 reference genome (mm10), and post-processing of the aligned data. In the second step, the combined aligned data is used to identify SNPs, indels, and structural variations, which are further quality filtered and functionally annotated.

Detection and distribution of SNPs and indels

We generated a catalogue of over 5.4 m high confidence sequence variants (**Figure 2-3**) including 4.46 m single nucleotide polymorphisms (SNPs) and 0.94 m insertions and deletions (indels). On average, there is one sequence variant present for every 500 bases in the reference genome. The ratio of transitions to transversions is 2.1:1 (3.02 m/1.44 m) which is strikingly similar to that observed in recent human studies particularly from the 1000 Genomes Project [2]. This indicates that majority of genetic variation between the two strains are spontaneous and neutral in nature. Indels are distributed evenly with respect to the parental genomes and the ratio of insertions (0.45 m) to deletions (0.49 m) is close to 1.

We assigned functional consequences to the sequence variants against RefSeq gene models. As expected, over sixty-percent (3.25 m) of variants including SNPs and indels are located in intergenic regions. A total of 1.77 m SNPs are located within genes, including introns (92.00%), exons (2.22%), 3' UTRs (1.74%), and 5' UTRs (0.30%). Approximately 32k and 34k SNPs are located within 2Kb of flanking region upstream and downstream of genes. Similarly, a total of 0.38 m indels are located within genes, including introns (93.20%), exons (0.39%), 3' UTRs (2.03%) and 5' UTRs (0.22%). 7.8k and 8k indels are located within upstream and downstream regions. The ~0.45 m insertions range in size from 1 bp to 33 bp and ~0.49 million deletions range in size from 1 bp to 55 bp. As expected, small indels are more frequent than large indels. The majority of small indels–65 % of the all insertions and 70% of all deletions—are 1 to 3 bp in length. Point indels account for 45% of insertions and 41% of deletions (**Figure 2-4**).

Variant density varies greatly across the genome (**Figure 2-3**). For example, a distal region of chromosome 1 from 170 Mb to 176 Mb known to be enriched for a number of behavioral and expression quantitative trait loci (QTLs) [187] has a high divergence rate (~5000 variants/Mb), whereas a 35 Mb region from 31 Mb to 66 Mb in the middle of chromosome 10 has a very low divergence (~30 variants/Mb). The unevenness in variant distribution can be attributed to complex history of the laboratory strains and the retention of long intervals that are almost identical by descent [188,189].

Functional consequences of coding SNPs and indels

Missense variants. We identified 39,268 exonic SNPs including 32,871 SNPs that are located in protein-coding genes. Of 32,871 coding SNPs, 21,976 are synonymous whereas 10,895 are nonsynonymous (missense, **Supplementary Table 2-1**). These 10,895 missense SNPs are located in 4,271 protein-coding genes. A total of 2,073 amino-acid substitutions (19.0% of missense SNPs) have potentially deleterious effects on function of 1,401 proteins based on analysis by PolyPhen2. Similarly, 1,760 substitutions (16.15%) have potentially deleterious effects on function of 1,275 proteins based on Grantham matrix scores. Five hundred and eighty five substitutions in 502 proteins are jointly identified to be deleterious. Proteins in the joint set are enriched for gene

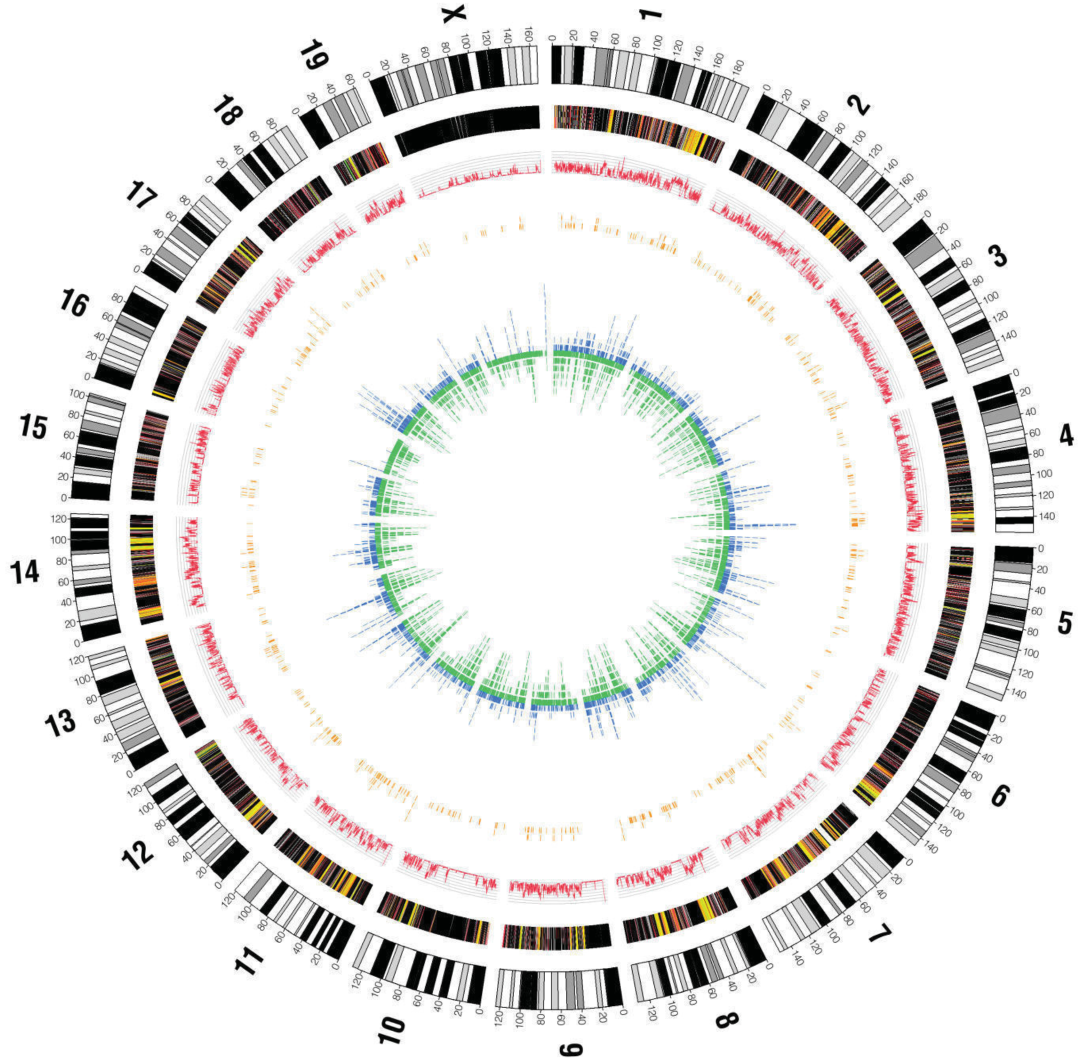


Figure 2-3. Distribution and density of sequence and structural variants along the DBA/2J genome

Notes. Moving inward from the outer circle, arcs in circle 1 denote chromosomes. Circle 2, SNP density with 100kb window (black is the lowest density and orange is the highest density). Circle 3, Indels density with 100kb window. Circle 4, Structural variants. Circle 5, CNVs, blue (outward) denotes loss of CNVs and green (inward) denotes gain of CNVs.

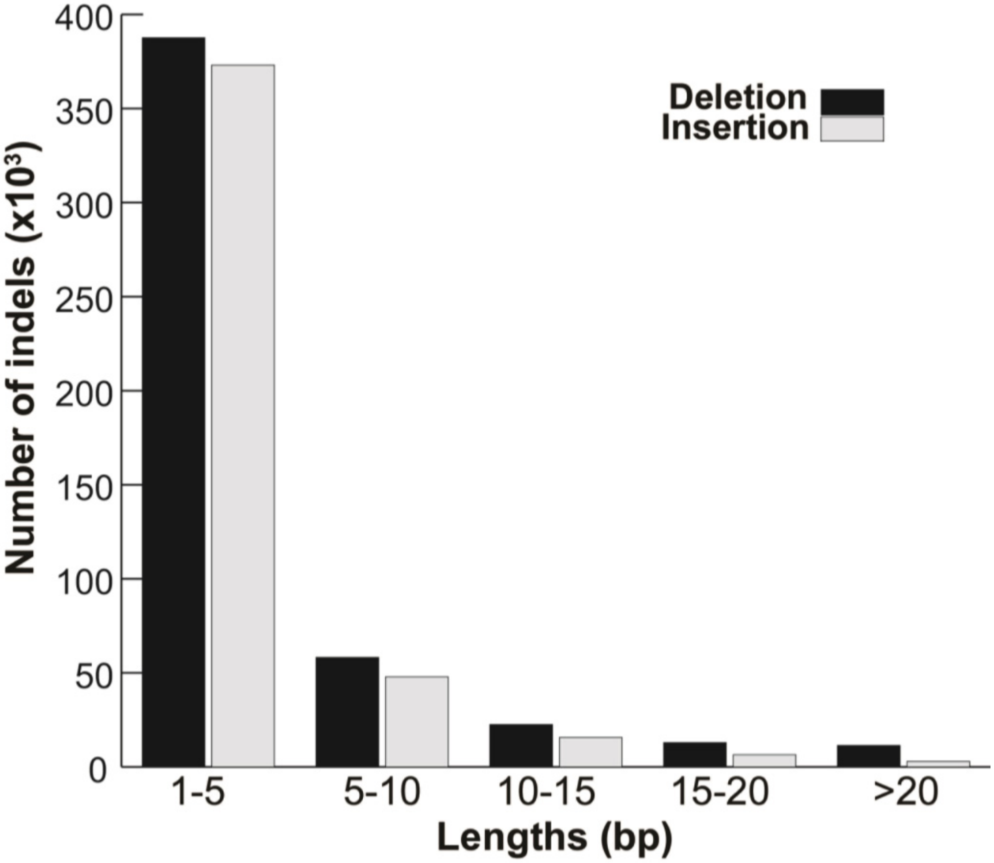


Figure 2-4. Distribution of indel lengths

Notes. The *X*-axis represents size of the indel and the *Y*-axis represents number of indels in thousands. Indels only up to 50 bp in length were considered for this analysis. As expected, the frequency of indels decreases with length.

ontology (GO) terms including sensory perception of chemical stimulus, olfactory receptor activity and neurological system process (Benjamini $p < 10^{-5}$). A large subset (20%) of the joint set includes olfactory, taste, and vomeronasal receptors.

Nonsense variants. We identified 66 SNPs, including 50 stop gains and 16 stop losses that affect a total of 63 genes. On manual inspection, we excluded five of these SNPs as nonsense variants due to ambiguous gene models and annotation error. Therefore, the final number of nonsense SNPs is 61, including 47 stop gains and 14 stop losses that affect a total of 58 genes (**Supplementary Table 2-1**). Forty-percent (*n*=25) of nonsense mutations result in protein variants with less than 10 % length differences compared to the reference protein. Eleven only differ by one to four amino acids in length. We also predicted that 33 nonsense mutations would produce truncated transcripts and 28 nonsense mutations would result in nonsense-mediated decay of the variant transcripts. We detected the two known nonsense mutations in *Gpnmb* [144] and *Ahr* [190]. DBA/2J is the only mouse strain out of 29 mouse strains (Mouse Genomes Project [172]) that has acquired a premature stop codon (**C**GA to **T**GA, *rs47598337*) in exon 4 of thirteen exons long *Gpnmb* resulting in a truncated transcript that presumably undergoes nonsense mediated decay. Expression variation in *Gpnmb* maps significantly to the location of the gene itself (**Figure 2-5a**). The nonsense variant in *Gpnmb* is a major contributing cause of pigment dispersion type glaucoma in DBA/2J [144].

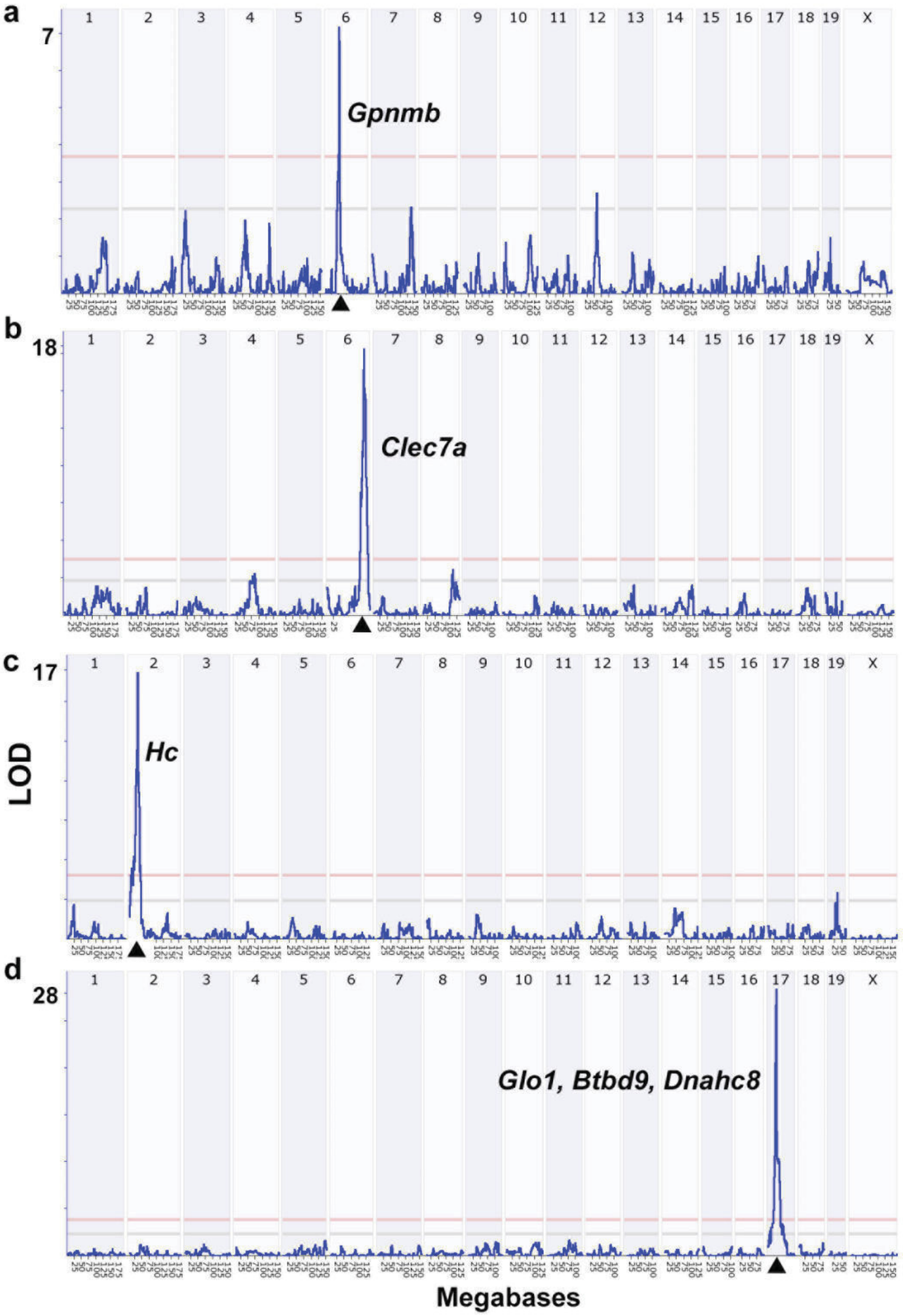
Splice-site variants. We identified 242 SNPs and indels that changed conserved bases at splice sites. We manually examined these variants and affected genes, and excluded 46 of these variants due to non-coding gene types and incorrect gene models. The final number of splice-site variants is 196 affecting 191 genes (**Supplementary Table 2-1**). An example is a **C** to **T** mutation (*rs30117984*) in the acceptor site of C-type lectin domain family 7 member A gene (*Clec7a*). Expression variation in *Clec7a* maps significantly to the location of the gene itself (**Figure 2-5b**). This splice site variant has been associated with susceptibility to infection with Coccidioides species [191].

Frameshifts and inframe indels. 735 small insertions and 770 small deletions are located within exons. Of these, 453 including 209 insertions and 244 deletions are located in protein-coding genes. Indels in coding sequence can be highly disruptive, especially when they introduce frameshift mutations resulting in abnormally short or abnormally long altered polypeptides. A subset of 99 small coding indels including 44 insertions and 55 deletions in 98 genes are predicted to result in frame shift mutations (**Supplementary Table 2-1**). An example is 'TA' deletion (*rs241579076*) in the hemolytic complement gene (*Hc*; chromosome 2 at 35.043208 Mb), that is associated with susceptibility to allergen- induced bronchial hyper-responsiveness [192] and intensified neurodegenerative responses [193]. Expression variation in *Hc* maps significantly to the location of the gene itself (**Figure 2-5c**).

The small coding indels are enriched in trinucleotides (multiples of three), which

Figure 2-5. Linkage maps of eQTLs for DBA/2J variants

Notes. **(a)** *Gpnmb* expression in eye is controlled by a *cis* eQTL with a peak LOD score of 7 located on Chr 6 at 48.9 Mb. This location corresponds precisely to the location of the *Gpnmb* gene (solid triangle). The horizontal lines provide genome-wide significance thresholds for the QTL determined by permutation analysis (upper <.05 and lower <.63) **(b)** *Clec7a* expression in spleen is controlled by a *cis* eQTL with a peak LOD score of 17.89 located on Chr 6 at 128.46 Mb **(c)** *Hc* expression in lung is controlled by a *cis* eQTL with a peak LOD score of 16.91 located on Chr 2 at 33.30 Mb **(d)** Combined eQTL mapping of *Glo1*, *Btbd9* and *Dnahc8* mRNAs in hippocampus. The eigenvalues associated with the first principal component maps with a LOD of ~28 to a CNV gain region spanning the above three genes in DBA/2J genome. Expression data used here were generated in GeneNetwork (www.genenetwork.org) using **(a)** BXD Eye M430v2 (Sep08) RMA Database, **(b)** UTHSC Affy MoGene 1.0 ST Spleen (Dec10) RMA Exon Level Database, **(c)** HZI Lung M430v2 (Apr08) RMA Database and **(d)** UMUTAffy Hippocampus Exon (Feb09).


a
7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Gpnmb
b
18
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Clec7a
c
17
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Hc
LOD
d
28
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Glo1, Btbd9, Dnahc8
Megabases

account for 78% (n=354) of total exonic indels including 165 insertions and 189 deletions (**Figure 2-6**, **Supplementary Table 2-1**). The trinucleotide indels are functionally not as severe as frameshift variants but they can still have a high functional impact on the structure and function of the protein. An example is an inframe deletion (*rs220745914*, chr2:102.901277 Mb) of 'GATGTG' nucleotides in the *Cd44* gene that deletes two amino acids.

Functional consequences of non-coding SNPs and indels

The vast majority of sequence variants are non-exonic and a particularly interesting subset of these non-coding variants may modulate transcription [194]. Sequence variants in the *cis* regulatory elements, including transcription factor binding sites, enhancers and 3' UTR microRNA binding sites affect gene expression. Recently, the mouse Encyclopedia of DNA Elements (ENCODE) consortium (http://www.nature.com/encode website) generated and published a plethora of sequence data that explores the genomic landscape of such elements in diverse mouse tissues [195]. Over 80% of 5.4 m variants had no potential functional impact based on mouse ENCODE data, but the remaining 20% (0.93 m) variants potentially affect one or more *cis* acting elements. Around 3% of them affect DNase I hypersensitive regions that mark open chromatin. ~15% of variants affect histone modifications including H3K36me3, H3K4me1, H3K4me2, H3K4me3 (active enhancers) and H3K27me3, H3K9me3, and H4K20me3 (inactive or poised enhancer sites). We also scanned 3' UTRs of mRNAs for variants in microRNA target sites. microRNA play a vital role in the regulation of gene expression by either translation inhibition or transcript degradation [196]. Three hundred and fifty variants disrupt 284 microRNA binding or target sites. Additionally, 36 variants are located within 31 mature microRNA coding sites in the microRNA primary transcripts.

Functional consequences of large structural variants

To identify structural variants, we used a combination of three different approaches including anomalous read-pairs, split-read and non-uniform read depth.

Split read approach. Pindel exploits reads with split alignment to identify exact breakpoints of the SVs. We detected ~2,919 insertions and ~14,486 deletions. The insertions range from over 50 bp to 31 kb and deletions range from over 50 bp to 32 kb. One hundred and eighty three indels affect exons or UTRs in 148 genes (**Supplementary Table 2-2**). An example is a 230 nucleotides deletion (chr16:18.407299-18.407528) in the 3' UTR of catechol-O-methyltransferase (*Comt*). This deletion variant has been associated with differences in the expression of genes involved in glutamatergic and GABAergic systems [88]. Additionally, it has also been associated with higher order phenotypes including sensitivity to dopamine receptors antagonist haloperidol and chloradiazepoxide, an allosteric modulator of GABA-A receptors [88].

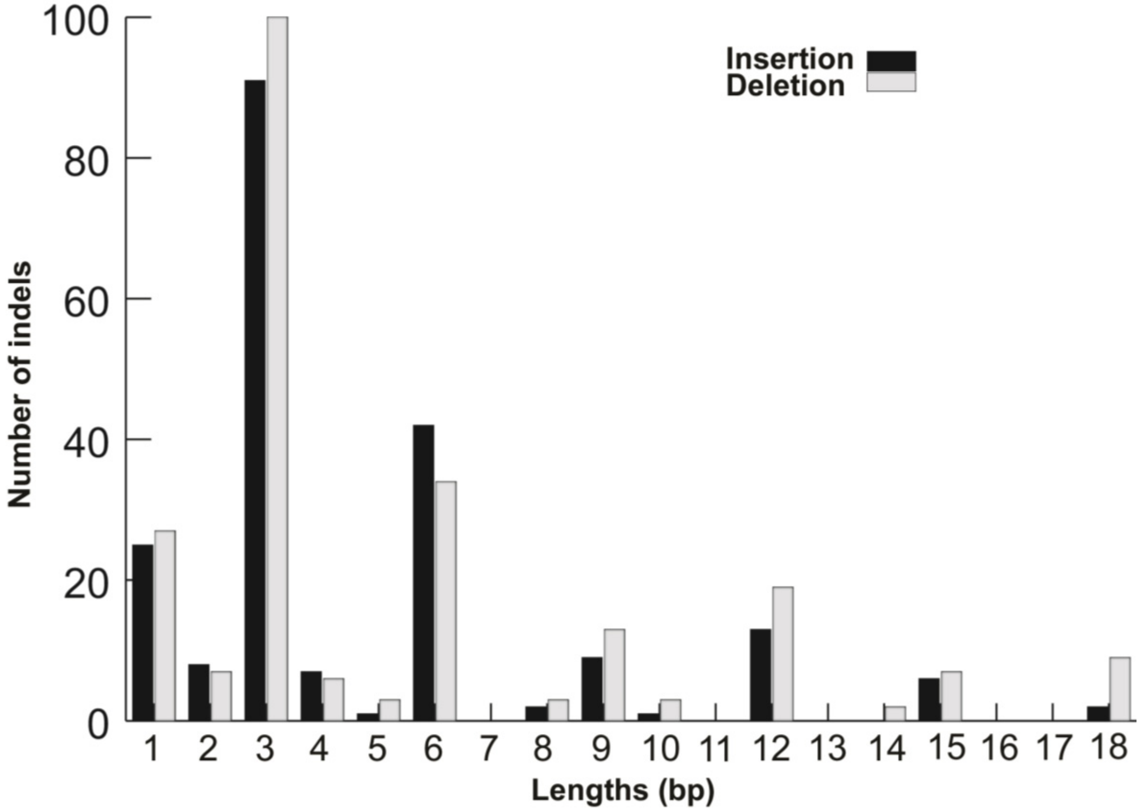


Figure 2-6. Distribution of lengths of coding indels

Notes. The *X*-axis represents size of the indel and the *Y*-axis represents number of indels. As expected, the coding indels are enriched in trinucleotides (multiples of three) that don’t result in frameshift variants.

Anomalous read-pairs. BreakDancerMax exploits read pairs with highly deviant insert sizes to identify SVs. We detected ~5,147 insertions and 28,025 deletions. The insertions range from over 100 bp to 0.5 kb and deletions range from over 100 bp to 976 kb. Five hundred and one indels affect exons or UTRs of 340 genes (**Supplementary Table 2-2**). An example is 0.66 Mb deletion (chr6:89.678881-90.331580) of large set of 16 vomeronasal receptors (V1r) in the DBA/2J genome. Deficient pheromone responses in mice have been associated with a significant reduction of male to male and maternal aggressiveness compared to controls [197].

Read depth. CND exploits genomic regions with highly deviant read depth or coverage to identify copy number variants (CNVs). We detected 7,615 CNVs, including 6,189 gains and 1,426 losses. Gains and losses have an average length of 3.3 kb and 4.3 kb respectively. Gains range from over 2 kb to 177 kb, and losses range from over 2 kb to 343 kb. A vast majority of copy number variants (~70%) affect intergenic regions. Of copy number gains relative to the C57BL/6J genome, 157 cover 215 genes completely (**Supplementary Table 2-2**). Of the losses, 40 cover 54 genes completely. An example is a copy number gain variant (chr17:30.443984-30.928714 Mb) that duplicates three genes entirely, including glyxolase 1 (*Glo1*), BTB domain containing 9 (*Btbd9*), and dynein axonemal heavy chain 8 (*Dnahc8*) in DBA/2J. Expression variation in these genes maps significantly to the location of the CNV (**Figure 2-5d**). The increased expression of *Glo1* in brain due to the CNV has been associated with anxiety-like behavior in mice [198,199]. GO enrichment analysis of the 269 that have been deleted or duplicated in their entirety revealed a significant enrichment (Benjamini-Hochberg *P* <0.01) of genes associated with 'MHC protein complex' and 'immune response'. A large subset (~20%) of these genes includes major histocompatibility complex genes (H2), killer cell lectin-like receptors (*Klra*), and genes associated with sensory perception including olfactory and vomeronasal receptors.

DBA/2J private variants

To obtain a list of private DBA/2J SNPs and indels, we compared DBA2/J variants with variant calls from 17 diverse strains of mice that were sequenced as a part of the Mouse Genome Project [172]. We classify a variant as private to DBA/2J only if all the other strains carry homozygous reference alleles at the variant site. We identified 69,554 private SNPs and 32,176 private indels in DBA/2J. They contribute to 1.5% and 3.42% of the total number of SNPs and indels in DBA/2J. Nearly half of the private SNPs (49%) and indels (46%) are located within intergenic regions. Three hundred and seventy three of the private SNPs result in non-synonymous mutations in 255 genes. Six private SNPs result in premature stop codon including *2310035K24Rik*, *Gm5592*, *Gpnmb*, *Klk1b8*, *Nbeal2,* and *Zfp277*. Six private indels introduce frameshift variants including *4930523C07Rik*, *Adam33*, *Caln1*, *Kif17*, *Tgif2,* and *Zfp354a.*

False positive rate detection

We resequenced 30 nonsense variants and 62 randomly selected missense variants using traditional Sanger sequencing to assess the false positive rates (FPR) of variants. We validated 27 nonsense variants and 60 missense variants—an FPR of 10% and 3.23% respectively. We also resequenced 20 splice-site variants and all of them were confirmed as true variants. We resequenced 13 frameshift variants and 11 of them were validated, indicating a false positive rate of 15.38%. Finally, we performed PCR-based validation of 40 predicted large structural variants. A total of 32 indels were validated, a ~20% false positive rate for detecting large indels.

DISCUSSION

The goal of this work has been to identify sequence and structural variants between the parental strains of the BXD family. To achieve this goal, we analyzed the DBA/2J genome at ~90X coverage using Illumina sequencing platform. We identified ~4.46 million SNPs, ~0.94 million indels, and ~20,000 high confidence SVs. The comprehensive catalog of sequence and structural variants that we uncovered provides an unprecedented resource with which to study the functional impact of naturally occurring sequence variants in a remarkably large and stable set of BXD lines. An interesting subset of high-impact functional variants can be utilized to initiate genome-wide reverse genetic analysis of complex traits such as PheWAS. Variant calls including SNPs, indels and SVs are available via a mirror of the UCSC genome browser (*ucscbrowserbeta.genenetwork.org*).

The genome of the DBA/2J strain has already been sequenced twice at lower coverage, initially by Celera Genomics [200] at about 1.3X using conventional Sanger sequencing, and by Mouse Genomes project (MGP) [172] at about 23X using almost precisely the same short read technology we have exploited. For this study, we combined sequence data from the MGP with our short-read data to enhance the sequence coverage. We compared high-confidence variant calls from our analysis with DBA/2J variant calls from the MGP [172]. We found an overlap of 4,621,903 variants including 4,056,910 SNPs and 564,993 indels (**Figure 2-7**). Of non-overlapping variants 941,379 and 487,796 were exclusively identified by our analysis and MGP respectively. We recalled 90% of SNPs and 87% of the indel calls from the MGP. We also added 469,212 novel SNPs including 1,297 non-synonymous SNPs and 472,167 novel indels. Our analysis substantially increased the number of known indels from 0.63 m to 1.1 m. This was expected as we used higher sequence coverage (100x vs. 50x) and longer reads (100 nt vs. 50 nt) compared to the MGP project. Longer reads generate more variant calls with higher acuracy [201] due to their higher alignability and mapping quality especially in low complexity regions. Longer reads also allow bigger gaps during alignments and are therefore more efficient in detecting relatively larger indels. Similar observations were made when raw (unfiltered) variant calls including high and low confidence variants were compared.

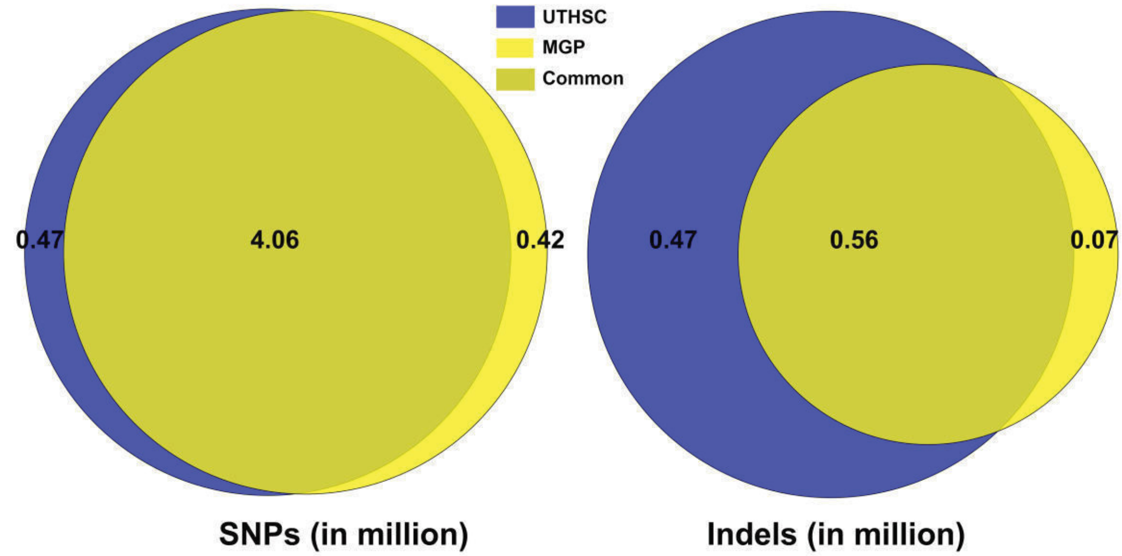


Figure 2-7. Comparison of DBA/2J variant calls between our data (UTHSC) and MGP

Note. The left and the right Venn diagrams compare filtered SNPs and indels respectively.

CHAPTER 3. GENOMIC ANALYSIS OF ALLELE-SPECIFIC EXPRESSION IN THE MOUSE LIVER

SYNOPSIS

Genetic differences in gene expression contribute significantly to phenotypic diversity and differences in disease susceptibility. In fact, the great majority of causal variants highlighted by genome-wide association are in non-coding regions that modulate expression. In order to quantify the extent of allelic differences in expression, we analyzed liver transcriptomes of isogenic F1 hybrid mice. Allele-specific expression (ASE) effects are pervasive and are detected in over 50% of assayed genes. Genes with strong ASE do not differ from those with no ASE with respect to their length or promoter complexity. However, they have a higher density of sequence variants, higher functional redundancy, and lower evolutionary conservation compared to genes with no ASE. Fifty percent of genes with no ASE are categorized as house-keeping genes. In contrast, the high ASE set may be critical in phenotype canalization. There is significant overlap between genes that exhibit ASE and those that exhibit strong *cis* expression quantitative trait loci (*cis* eQTLs) identified using large genetic expression data sets. Eighty percent of genes with *cis* eQTLs also have strong ASE effects. Conversely, 40% of genes with ASE effects are associated with strong *cis* eQTLs. *Cis*-acting variation detected at the protein level is also detected at the transcript level, but the converse is not true. ASE is a highly sensitive and direct method to quantify *cis*-acting variation in gene expression and complements and extends classic *cis* eQTL analysis. ASE differences can be combined with coding variants to produce a key resource of functional variants for precision medicine and genome-to-phenome mapping.

INTRODUCTION

Genetic variation contributes greatly to phenotypic diversity and differences in disease susceptibility by altering the structure and expression levels of proteins. The analysis of complex phenotypes in the pre-genomic era focused on coding variants, especially including nonsense, missense, and frameshift mutations. However genome-wide association studies conducted over the last decade have demonstrated that a great majority (>90%) of trait/disease-associated variants are located in non-coding regions. These non-coding variants primarily act by modulating gene expression, and they are the major cause of variation in susceptibility to complex diseases [94-96].

Sequence variants that affect gene expression can act in *cis* or in *trans*. *Cis*-acting variants represent first-order local control of gene expression that is specific to each individual haplotype. For example, sequence variants in transcription factor binding sites may affect expression of cognate genes on the same chromosome. Cis-acting variants are key to understanding heritable variation in disease risk, and serve as direct targets for diagnosis and treatment of diseases. Cis-acting variants can, of course, also have second-order distal or *trans* effects. A small subset of cis-modulated transcripts consists of master trans-regulators (for example, transcription factors, miRNAs) that control the abundance of large numbers of downstream target genes on both sets of chromosomes. Hence, identification of cis-modulated transcripts serves as an key molecular resource for reverse genetics studies that focus on downstream consequences of altered expression.

Currently, two genome-wide approaches can be employed to identify cis-acting variation in expression. The first approach, known as expression quantitative trait locus (eQTL) mapping, performs classical genetic linkage analysis of expression usually for an entire transcriptome or proteome. This approach has been widely applied to study the effects of segregating variation on gene expression in yeast, mice, maize, and humans [97,142,202-204]. The largest study to date is the ongoing Genotype-Tissue Expression (GTEx) project that is generating a comprehensive resource of *cis* eQTLs for multiple tissues in a large human cohort [205,206]. The second approach, widely used in studies of model organisms, exploits rtPCR or RNA-seq to assay allele-specific expression (ASE) differences in isogenic heterozygous (F1) individuals [207-211]. RNA-seq can reliably distinguish mRNAs transcribed from the alternative alleles, and can be used to detect unequal production of the two alleles. A major advantage of isogenic F1 hybrids is that they provide a way to control for environmental and trans-acting influences. Both alleles are present within an identical environment and subjected to the same genetic background and regulatory networks. As a result, any expression differences between alleles in an isogenic F1 can be confidently attributed to genetic or epigenetic regulatory variant acting in *cis* [212-214].

In this study we evaluate and compare the impact of cis-acting variation on expression in murine liver using both ASE and eQTL approaches. We exploit RNA-seq data from isogenic F1 hybrids and array data from a large set of recombinant inbred strains of mice—the BXD cohort, and generate a molecular resource for genome-wide reverse genetics that focuses on downstream consequences of altered gene expression [88,171]. We address the following questions:

- How do genes that exhibit ASE differ from those that do not?
- How do the two approaches highlighted above compare in terms of detecting effects of local polymorphisms on expression? More specifically, are cis-modulated transcripts identified by eQTL mapping also consistently detected by ASE analysis?
- How frequently do cis-acting variants that cause mRNA differences also cause differences in protein expression?

MATERIALS AND METHODS

RNAseq data for C57BL/6JxDBA/2J hybrids

We downloaded paired-end RNA-seq data from the European Nucleotide Archive (accession number ERP000591) for liver of C57BL/6JxDBA/2J F1 female hybrids generated by crossing C57BL/6J females with DBA/2J males [172]. The data consist of transcriptome sequence from six biological replicates. We acquired a total of ~181 m read pairs (2x76 nt in length). We removed low quality reads and used the remaining ~173 m read pairs for alignment.

RNA-seq read alignment

We aligned RNA-seq reads to both the C57BL/6J reference genome (mm10 assembly) and the DBA/2J genome using "Splice Transcripts Alignment to a Reference" tool (STAR, version 2.3.1a) [215] with the following parameters "--outFilterMultimapNmax 10 --outFilterMismatchNmax 12". Read pairs that were not aligned in concordance with the library design, in particular read strand, were removed. We allowed a maximum insert size of 300,000 nucleotides (maximum intron length) to allow alignment of those read-pairs aligned to different exons. We selected read pairs for which both reads were uniquely aligned and for which each had less than six mismatches. If one member of a read-pair could not be aligned then we retained the other member only if it could be aligned uniquely.

Calculation of allelic ratio

We used SAMtools (version 0.1.19) "pileup" function [176] and an in-house Python (https://github.com/ashutoshkpandey/ASE_prealignment/blob/master/Allele_specific_SAM.py) script to assign reads to their parental allelic origin by comparing alignments to the C57BL/6J and the SNP-substituted DBA/2J genome. If reads were aligned to both genomes then we required them to map at the same locations. Those reads that overlapped SNPs were assigned to their parental allele origin. To ensure that differential expression was not due to amplification by PCR during library preparation, we removed all potential PCR duplicates except for the single read with the fewest mismatches using Picard's MarkDuplicates tool (version 1.78). We calculated allelic ratios for each SNP defined as the ratio of number of reads assigned to the reference allele (B) to the total number of aligned reads (B+D).

Definition of ASE using chi-square test

For each SNP we used an interquartile range (IQR) method to identify outlier allelic ratios from the set of F1 replicates. Outlier ratios were located outside the [Q1 – 1.5(IQR) and Q3 + 1.5(IQR)] range where Q1 and Q3 represent first and third quartiles and IQR is calculated as Q3 – Q1. Reads from replicates showing concordant allelic ratios were merged and allelic ratios were recalculated. We used the chi-square goodness of fit test to determine allelic imbalances for a given SNP. For a SNP showing an allelic imbalance, the ratio will deviate from 0.5. We defined genes as having an allele-specific expression difference if they contained one or more SNPs with an allelic imbalance at an

FDR threshold of less than 0.1 [216]. We also required the expression fold difference to be >1.25.

Array expression data and eQTL mapping

We used an Affymetrix data set (Mouse 430 v2.0 array) consisting of liver gene expression data for 40 genetically diverse BXD strains (GeneNetwork.org accession GN310, http://genenetwork.org/webqtl/main.py?FormID=sharinginfo&GN_AccessionId=310). We performed robust multichip analysis (RMA) preprocessing and rescaled values to $\log_2$ and stabilized the variance across samples [217]. We used QTL Reaper, mapping code that uses the method of Haley and Knott for eQTL analysis [218], and a set of 3,200 markers. We excluded probe sets located on X and Y chromosomes (~2,500 probe sets). Locations of probe sets were identified using custom annotation files. Similarly, we performed pQTL mapping on expression data from 172 proteins [219]. This data can be downloaded from Genenetwork.org (accession GN490, http://www.genenetwork.org/webqtl/main.py?FormID=sharinginfo&GN_AccessionId=490). To identify Affymetrix probes that overlapped sequence variants, we first aligned probe sequences against the mouse reference genome (mm10) using BLAT [220], and then compared genomic coordinates of probes for overlap with sequence variants.

Comparison between ASE and non-ASE genes (URLs)

We downloaded the liver-specific regulatory elements data from Ensembl Regulatory build (ftp://ftp.ensembl.org/pub/release-81/regulation/mus_musculus); see more details on this build here: http://www.ensembl.org/info/genome/funcgen/regulation_sources.html. For TFBS comparison we used data from MotifMap—genome-wide maps of regulatory elements. The file was downloaded using the following link: (http://www.igb.uci.edu/~motifmap/motifmap/MOUSE/mm9/multiz30way/MotifMap_MOUSE_mm9.multiz30way.tsv.bz2). A list of house-keeping genes was downloaded using the following link: http://www.tau.ac.il/~elieis/HKG. In order to compare for the evolutionary conservation between the ASE and non-ASE genes, we used GERP++ scores for mouse (http://mendel.stanford.edu/SidowLab/downloads/gerp/mm9.GERP_elements.tar.gz). We downloaded *M. musculus* and *H. sapiens* paralog data from Ensembl BioMart (www.ensembl.org/info/data/biomart.html) [221,222]. All the mm9 coordinates were converted to mm10 using UCSC liftOver utility. The counts/scores of cis-regulatory elements, TFBSs, DBA/2J sequence variants, and GERP++ scores were normalized by gene length before comparison.

Single marker analysis

We performed single marker analysis as an alternative to eQTL mapping to identify cis-modulation in expression. For each gene we selected its closest marker and classified BXDs by genotype (*B* allele or *D* allele) for that marker. We compared expression using a *t*-test and selected genes showing significant expression difference at an FDR of < 0.1.

RESULTS

DBA/2J specific reference genome

We substituted ~4.5 million DBA/2J SNPs (Chapter 1) into the reference genome (GRCm38/mm10) to create a customized DBA/2J genome for RNA-seq read alignment. A total of ~1.7 m SNPs are located within coding genes based on RefSeq annotation, including introns (95.59%), exons (2.30%), 3' UTRs (1.80%) and 5' UTRs (0.30%). These SNPs are distributed among 14,591 genes. SNPs in transcribed regions were used to discriminate between, and identify, the parental allelic origin (*B* vs *D*) of transcripts in isogenic F1 hybrids.

Haplotype-aware alignment corrects for allelic bias in read alignment

We downloaded paired-end liver RNA-seq reads for six biological replicates of C57BL/6JxDBA/2J F1 females. We adopted a haplotype-aware alignment approach and aligned ~350 m (~175 m paired-end) reads against both the *B* and the customized *D* genomes (Methods). We used a SNP-directed approach to identify the allelic origin (*B* or *D*) of reads that aligned over heterozygous SNPs in the F1 samples. Only uniquely aligned reads were assigned to parental alleles. Approximately 0.27 m SNPs within genes (a great majority within exons) had at least one read.

RNA-seq read alignment suffers from allelic bias that disfavors reads containing sequence variants relative to the reference genome [223,224]. This bias generates lower read counts for non-reference alleles, and overestimates ASE differences. To evaluate bias, we examined allelic ratios—defined as the number of reads with the reference allele (*B*) divided by the total number of reads (*B*+*D*). In the absence of bias, this ratio will have a symmetrical distribution and a mean of 0.5. For each of the F1 samples, ratios were well balanced, with nearly equal numbers of SNPs with high *B* or high *D* expression. Additionally, mean and median ratios were close to 0.5 indicating that the majority of SNPs exhibit small or undetectable ASE. We compared our results with a traditional approach involving alignment of reads against the reference genome and allowing for fewer mismatches (1 mismatch per 25 nt). This produced an artifactually high number of SNPs with high *B* expression (~ 3,000 *B* vs ~400 *D*, two-sided binomial $p < 10^{-323}$) compared to the dual genome alignment (~2,325 *B* vs ~ 2,300 *D*, two-sided binomial *p* value = 0.724, **Figure 3-1a**, **b**). The mean and median of allelic ratios using the standard approach were also skewed—0.69 (high *B*) and 0.68, respectively. This illustrates that the haplotype-aware alignment workflow is highly effective in reducing allelic bias due to read alignment.

High correlation of allelic ratios across biological replicates

We calculated the correlation of allelic ratios with read depth ≥ 20 across all biological replicates. Allelic ratios were highly correlated with an average Pearson correlation of 0.70 ± 0.02 for all pairs of replicates (n = 15). We merged data from biological replicates, but to minimize variation across replicates, we discarded reads from replicate with highly discrepant ratio (Methods). More than 90% of SNPs had closely matched ratios across four or more replicates and were retained for further analyses. We also checked the concordance of the polarity of ASE measured by neighboring (<75 nucleotides) but independent SNPs. In the great majority of cases SNPs within the same genomic feature (5' UTR, exon, intron and 3' UTR) were highly concordant. Only 136 (6%) of 2,234 genomic features contained SNPs with opposite ASE polarity and the great majority were in 3' UTRs (n = 86). 3' UTRs undergo extensive alternative processing [225,226], and SNPs with opposite ASE polarity probably represent alternative polyadenylation sites [88]. The high correlation of allelic ratios across the replicates and the high concordance in polarity again demonstrate the accuracy of the haplotype-aware alignment workflow. SNPs located within copy number variants, large insertions and deletions or in close vicinity (< 75 nucleotides) to an indel can generate inaccurate ASE estimates due to alignment artifacts. As a result, these SNPs were removed from further analysis. Additionally, to ensure independent sampling we considered only one SNP of a SNP pair when SNPs were separated by less than 75 nucleotides. Of 21,166 SNPs with read coverage ≥ 30, ~25% (5,358) SNPs were removed for one of these reasons.

ASE differences in liver are common

We tested the null hypothesis of equal abundance of transcripts representing *B* and *D* alleles in isogenic F1 hybrids using a Chi-square Goodness of fit test (FDR < 0.1, Methods). On average we used ~650 reads per SNP to test for ASE. At a minimum threshold of 30 reads per SNP, we were able to test 15,808 SNPs in 3,589 genes (**Supplementary Table 3-1**). We detected significant ASE in 5,298 SNPs from 1,905 genes (**Supplementary Table 3-1**, **Figure 3-2**). Most of these SNPs are contained within coding exons (40%) and 3' UTRs 40%) (**Supplementary Table 3-1**). Seven percent are in introns and may represent unannotated exons or transcripts with unspliced or retained introns. We obtained comparable results when the minimum read threshold was increased to ≥ 60 (4,968 SNPs in 1,791 genes) and when the FDR threshold was decreased to 0.05 (4,774 significant SNPs in 1,482 genes). Fifty-two percent (2,773 *B* vs 2,525 *D*) of SNPs have higher expression from the *B* allele. There is no difference in the distribution of average effect sizes for significant ASE between alleles (**Figure 3-3**). Over half of the SNPs with significant ASE differ by less than two fold; one-third differ two–four folds;

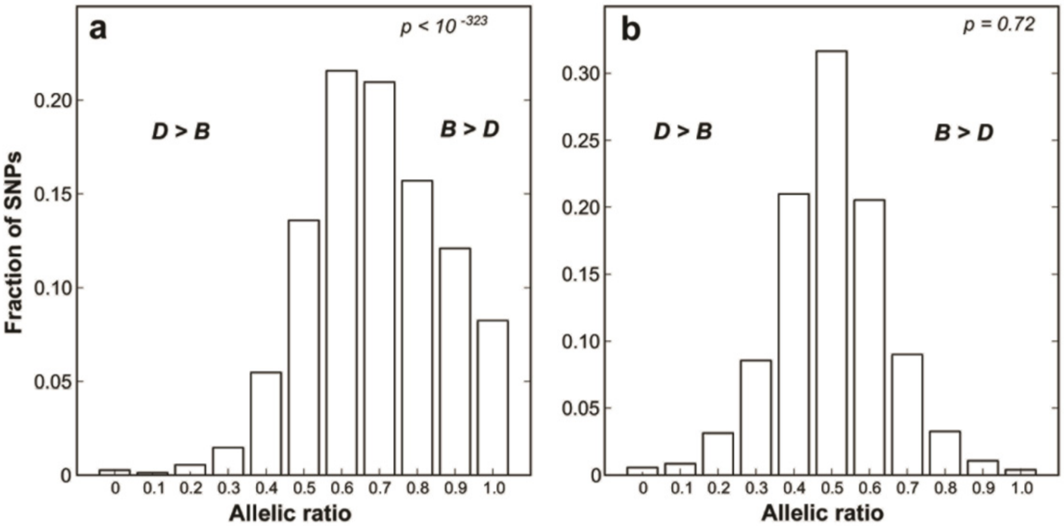


Figure 3-1. Comparison of the allelic bias in read alignment between traditional and haplotype-sensitive approach

Note. Distribution of allelic ratios in **(a)** traditional and **(b)** haplotype-sensitive alignment.

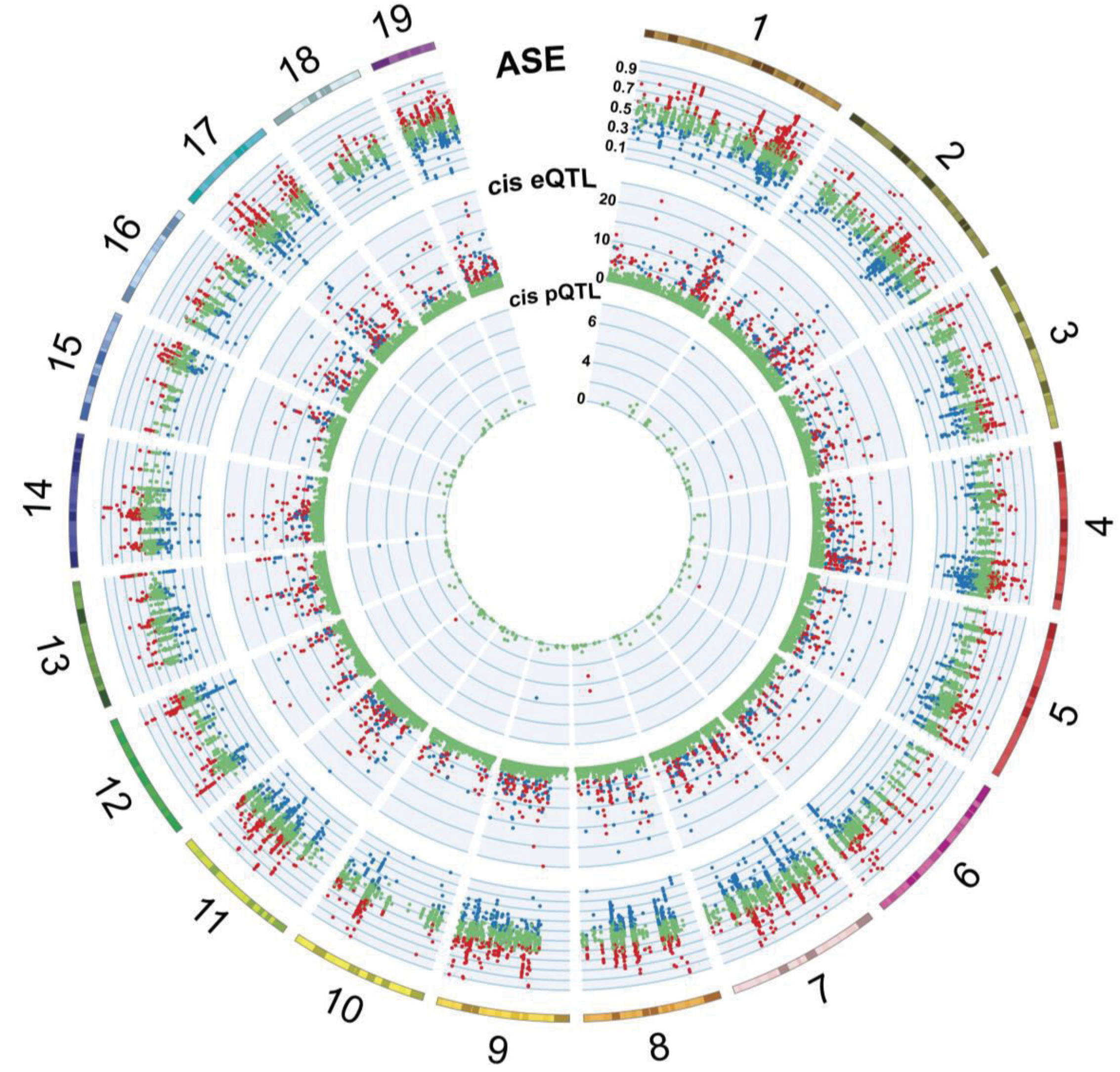


Figure 3-2. Distribution of cis-modulated genes in liver

Notes. The outermost circle represents chromosomes. Moving in, the second circle represents a scatter plot of ~15,000 SNPs tested for ASE. The *Y*-axis represents the allelic ratio. SNPs with significant ASE are shown in red and blue, representing high expression of the *B* and the *D* allele respectively. SNPs with insignificant ASE are shown in green. These SNPs are located on or near the line representing an allelic ratio of 0.5. The third circle represents a scatter plot of ~40,000 microarray probe sets tested for *cis* eQTLs. The *Y*-axis represents the LOD scores of probe sets measured at the nearest marker (*cis* LOD). *Cis* LOD (≥ 3) scores associated with high expression of the *B* and the *D* allele are shown in red and blue respectively. *Cis* LOD scores of less than 3 are shown in green. The innermost circle represents a scatter plot of ~200 proteins tested for *cis* pQTLs. The *Y*-axis represents the LOD scores of proteins measured at the nearest marker (*cis* LOD). Chromosomes X and Y were excluded.

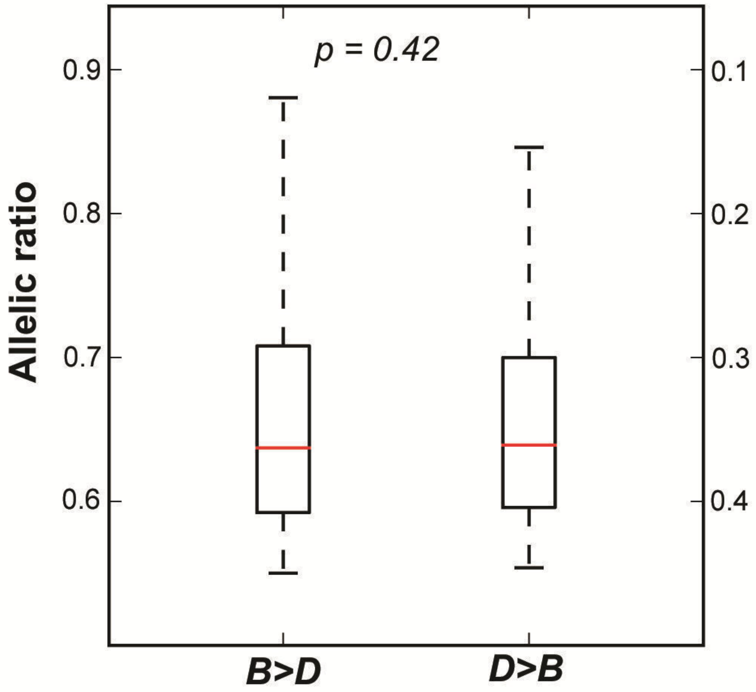


Figure 3-3. Distribution of significant allelic ratios

Notes. The left boxplot labelled as "*B>D*" represents SNPs with high expression of the *B* allele (left *Y*-axis). The right boxplot represents SNPs with high expression of the *D* allele (right *Y*-axis). The *Y*-axis represents allelic ratios. Outliers are not shown.

and the remaining one-sixth differ more than four-fold. We detected the known low *D* allele expression of aryl hydrocarbon receptor (*Ahr*)—a transcription factor that controls xenobiotic metabolizing enzymes such as cytochrome P450 gene family [227]. Similarly we also detected the known low *B* allele expression of alkaline phosphatase (*Alpl*)—a gene linked to hypophosphatasia [86].

Comparison between ASE and non-ASE genes

Genes with high or low levels of ASE may differ in length, complexity of promoters, sequence variant density, or evolutionary history. To explore these differences we selected a subset of 418 genes with very high ASE ratios (>1.5 fold) and a subset of 465 genes with low or no ASE. All genes in both groups were required to have at least two independent SNPs that supported their categorization. We also required all SNPs to have more than 100 supporting reads—roughly the top ten percentile. The average read depth for ASE and non ASE SNPs were 258 ± 180 and 262 ± 168. We defined each gene as the region between the transcription start site and 3' UTR with 2 kb of flanking regions upstream and downstream. ASE genes do not differ from non-ASE genes in terms of total gene length or their 5' or 3' UTR length (**Table 3-1**). They also do not differ in numbers of protein-coding transcripts (isoforms) or numbers of exons per transcript (**Table 3-1**). However, ASE genes have a higher functional redundancy (number of paralogs) compared to non-ASE genes (1.5 fold, $p < 10^{-4}$, **Table 3-1**).

We also compared promoter complexity. There are no differences in the density of liver-specific cis-regulatory elements defined using mouse ENCODE data [228]. Similarly, there are no differences in the density of transcription factor binding sites defined using a comparative genomic approach [229] (**Table 3-1**). However, the subset of genes with no or low ASE are enriched ($p < 10^{-46}$, hypergeometric test) in housekeeping genes [230]. In fact, nearly 50% of the non-ASE set are house-keeping genes. In contrast only 20% of the ASE set belong to this category.

Another distinguishing characteristic of the two sets is their density of sequence variants. The mean density in the non-ASE set is significantly different from the ASE set (4.59 ± 0.02 versus 8.20 ± 0.25 per Kb, p ~ 0.0, two-tailed *t* test). This suggests that genes in the non-ASE set are under comparatively stronger purifying selection.

To test whether ASE and non-ASE sets are subject to different levels of purifying selection (the elimination of deleterious sequence variants) we compared the strength of selective constraint (GERP++ scores) on genomic regions across 33 mammalian species [125,126]. The ASE gene set (201.33 ± 10.85) have significantly lower conservation scores than the non-ASE set (274.15 ± 13.26) indicating that they tolerate and accumulate more mutations; a subset of which are highly likely to modulate expression ($p < 10^{-4}$, two-tailed *t* test, **Table 3-1**).

To evaluate if genes in ASE and non-ASE sets belong to different functional categories, we compared them for overrepresented gene ontology and KEGG pathway

Table 3-1. Comparison between ASE and non ASE genes

Category	ASE	Non ASE	P-value
Gene length in Kilobase (Kb)	58 ± 40	63 ± 30	0.31
5’ UTR length (Kb)	0.23 ± 0.03	0.25 ± 0.01	0.15
3’ UTR length (Kb)	1.80 ± 0.08	1.87 ± 0.06	0.54
Coding transcripts per gene	1.50 ± 0.05	1.65 ± 0.06	0.1
Exons per transcript	12.40 ± 0.37	12.90 ± 0.29	0.31
Cis-regulatory elements (Kb)	0.37 ± 0.01	0.34 ± 0.01	0.11
Transcription factor binding sites (Kb)	1.61 ± 0.07	1.70 ± 0.07	0.38
Sequence variants (Kb)	8.20 ± 0.25	4.59 ± 0.02	0.00000001
Paralogs per gene	5.31 ± 0.36	3.56 ± 0.22	0.0000286
Conservation score (Kb)	201.33 ± 10.8	274.15 ± 13.2	0.0000331

terms using DAVID functional annotation tool [231]. ASE genes are significantly enriched (Benjamini corrected $p < 0.05$) in genes associated with KEGG pathway terms including ‘complement and coagulation cascades’, ‘retinol metabolism’, ‘metabolism of xenobiotics by cytochrome P450’, and ‘lipid, fatty acid and steroid metabolism’. Non-ASE genes are enriched in genes with gene ontology (GO) terms representing broad functional categories such as ‘macromolecule localization’, ‘catalytic activity’, and ‘ubiquitin mediated proteolysis’.

To evaluate the effect of ASE on phenotypes we performed phenotype enrichment analysis [232] on mouse-mutant phenotypes [233] derived from Mouse Genome Informatics (MGI, www.informatics.jax.org/phenotypes.shtml). As noted above, ASE genes compared to non-ASE genes are enriched (unadjusted $p < 0.01$) for phenotypes including ‘abnormal gall bladder physiology’ (MP:0005085), ‘abnormal xenobiotic induced morbidity/mortality’ (MP:0009765), and ‘abnormal glucose homeostasis’ (MP:0002078).

Identification of *cis* eQTLs

We performed linkage-based eQTL mapping using a gene expression data set generated using liver samples from 40 BXD strains (Methods). Of the ~45,000 probe sets, we selected ~41,500 that have a uniquely assigned gene identifier. This subset represents ~19,000 genes. *Cis* eQTLs were required to have LOD scores greater than 3 and LOD peaks within ± 5 Mb of their cognate gene. A LOD score ≥ 3 roughly corresponds to a nominal p value of < 0.001 and is widely used to indicate a high probability of linkage. We detected a total of 1,907 *cis* eQTLs corresponding to 1,474 genes (**Supplementary Table 3-2**). *cis* eQTLs with very high LOD scores (≥ 25) include *Snx6, Adi1, Cfh, Fbxo39,* and *St3gal4.*

Variant overlapping probes cause spurious *cis* eQTLs

SNPs and indels in probe sequences can influence hybridization kinetics and cause incorrect measurement of expression. Twenty-five percent of apparent *cis* eQTLs detected in the hippocampus are probably caused by variants in probes rather than by genuine differences in expression [207]. We identified 739 cis-modulated probe sets that overlap *D* variants. To evaluate how these variants affect the direction and size of additive effects for corresponding *cis* eQTLs, we compared *cis* eQTLs between probe sets with variants and probe sets without variants. Probe sets without variants were precisely balanced with respect to *B* versus *D* effects. In contrast probes with variants were highly imbalanced and ~70% were associated with high *B* expression. Of 408 *cis* eQTL genes represented by probes with variants, 193 could be compared with results from ASE analysis, and 149 genes showed the same direction of expression bias as ASE. A total of 1,215 genes were associated with *cis* eQTLs.

Cis-modulated genes from ASE and eQTL mapping overlap

We compared results from ASE with those from eQTL mapping. Of the 3,431 genes that were jointly tested, 1,808 (~50%) and 867 (~25%) were identified as cis-modulated by ASE and eQTL mapping, respectively. Six hundred and eighty-three genes were jointly identified as cis-modulated, a significant overlap (hypergeometric $p < 10^{-73}$), and ~90% had the same effect polarity. One thousand one hundred and twenty-five and 184 cis-modulated genes were exclusively identified by ASE and eQTL mapping respectively. In other words, roughly 80% of *cis* eQTLs also have ASE differences and ~40% of ASE differences are associated with *cis* eQTLs. To investigate discrepancies, we compared LOD scores of jointly identified cis-modulated genes with those only identified by eQTL mapping. The joint set exhibit significantly higher LOD scores ($p < 10^{-5}$) (**Figure 3-4**). We further compared the RNA-seq read depth (expression) of these two groups and the joint set has significantly higher expression (three-fold difference, $p = .03$, **Figure 3-5**). This suggests that the ASE analysis lacks adequate read depth (statistical power) to detect allelic differences corresponding to *cis* eQTLs with comparatively low LOD scores. We performed an empirical power analysis (**Figure 3-6**) to illustrate dependency of ASE analysis on read-depth to detect allelic differences of different magnitudes. As expected, strong differences can be reliably detected with a relatively small number of reads and vice-versa. The joint set also has higher allelic expression differences compared to 1,125 genes identified only by ASE ($p < 10^{-30}$) (**Figure 3-7**). We used a stringent LOD threshold ≥ 3 to define *cis* eQTLs and this will reduce the number of *cis* eQTLs corresponding to genes with low ASE. We therefore performed single marker analysis at less stringent FDR < 0.1 (Methods) to identify cis-modulated genes and compared them with the ASE gene set. The number of jointly identified genes increased from 683 (eQTL mapping, LOD ≥ 3) to 962 (single marker analysis, FDR < 0.1), and those exclusively identified by ASE were reduced from 1,125 to 774. Thus a large fraction of disjoint between ASE and eQTL results are explained by the different statistical criteria and thresholds we used to define both ASE genes and *cis* eQTLs.

Genetic variants affecting transcript abundance and protein abundance show poor overlap

We performed linkage based protein QTL (pQTL) mapping on liver proteomics data generated from a set of 38 BXD strains [219]. One hundred and seventy-two autosomal proteins involved in metabolism were quantified using a targeted mass spectrometry method. Only 7% ($n = 12$) are associated with *cis* pQTLs, including ABCB8, ACADS, ACOX1, ATP5O, BCKDHB, CAR3, DHTKD1, GCLM, MRI1, NNT, PM20D1, and TYMP (**Figure 3-8**, where a *cis* pQTL must have a LOD > 2 located within ±5 Mb of the parent gene). Not surprisingly, all of the *cis* pQTLs are also associated with significant ASE differences with matched polarity. Similarly, 8 of these *cis* pQTLs are linked to *cis* eQTLs with high LOD scores and with matched polarity. However, 39 genes with significant ASE and 18 genes with significant *cis* eQTL are not associated with *cis* pQTLs. For example, *Ddah1* has significant ASE

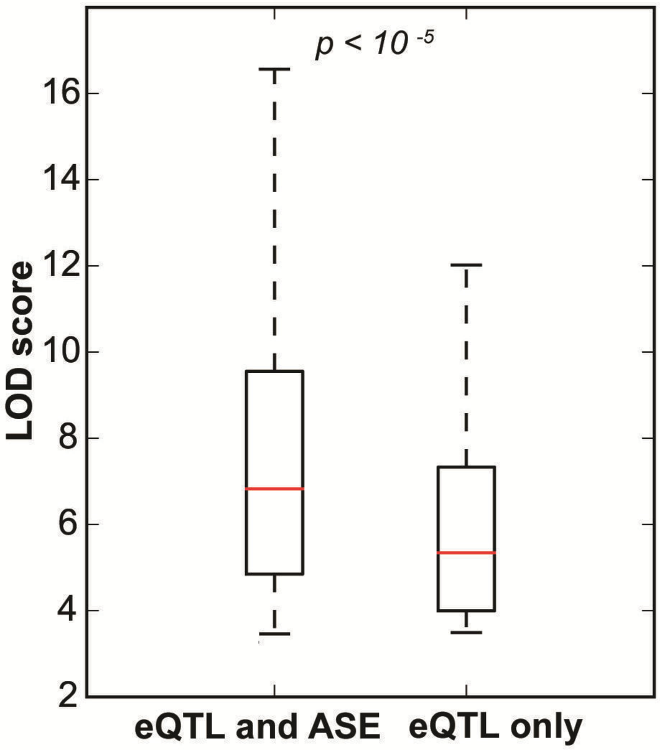


Figure 3-4. Comparison of LOD scores

Notes. Comparison of LOD scores from jointly identified cis-modulated genes (ASE and eQTL mapping, left boxplot) with those only identified using eQTL mapping (right boxplot). The *Y*-axis represents LOD scores. Outliers are not shown.

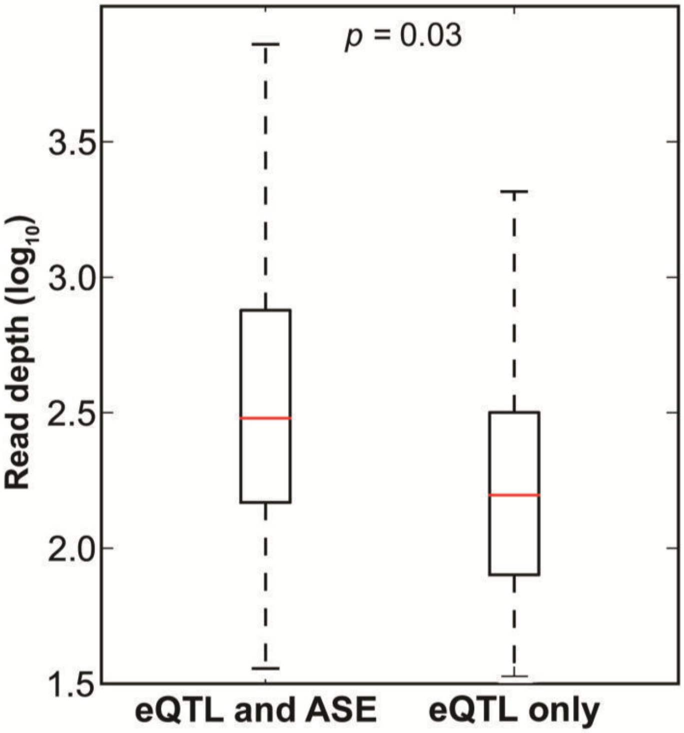


Figure 3-5. Comparison of RNA-seq read depth ($\log_{10}$)

Note. Comparison of RNA-seq read depth ($\log_{10}$) from jointly identified cis-modulated genes (ASE and eQTL mapping) with those only identified using eQTL.

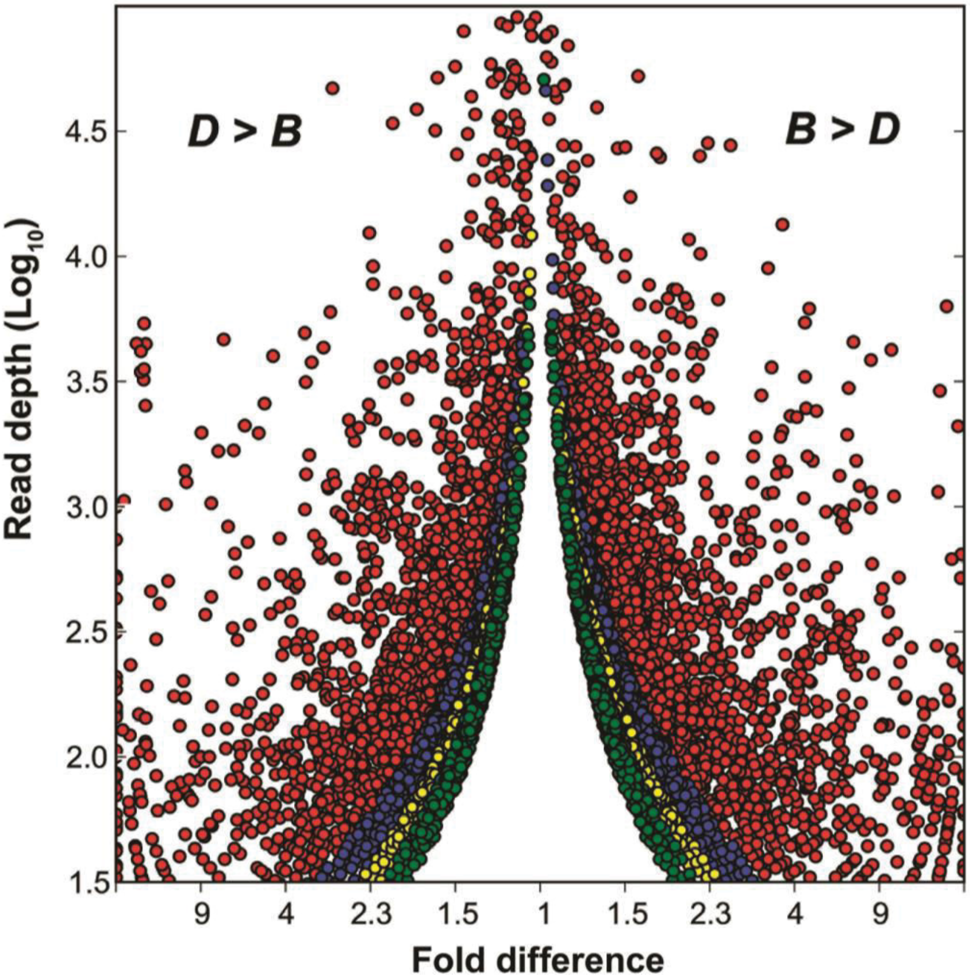


Figure 3-6. Empirical determination of read depth required to detect allelic differences of a given size

Notes: Each circle represents a SNP. The *X*-axis represents the measured fold-difference and the *Y*-axis represents RNA-seq read depth ($\log_{10}$) for a given SNP. SNPs exhibiting ASE at an FDR threshold of ≤ 0.2 have been plotted as circles. Red circles represent SNPs with ASE at an FDR threshold of ≤ 0.01. The red and blue circles, combined, represent SNPs with ASE at an FDR threshold of ≤ 0.05. Similarly, red, blue and yellow circles, combined, represent SNPs with ASE at an FDR threshold of ≤ 0.1.

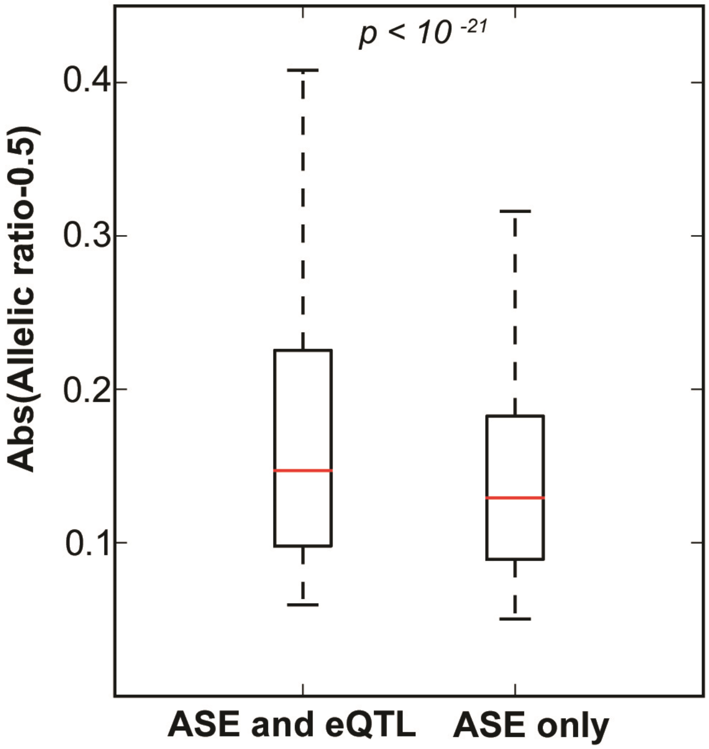


Figure 3-7. Comparison of absolute allelic differences

Notes. Comparison of absolute allelic differences from jointly identified cis-modulated genes (ASE and eQTL mapping, left boxplot) with those only identified using ASE (right boxplot). Outliers are not shown.

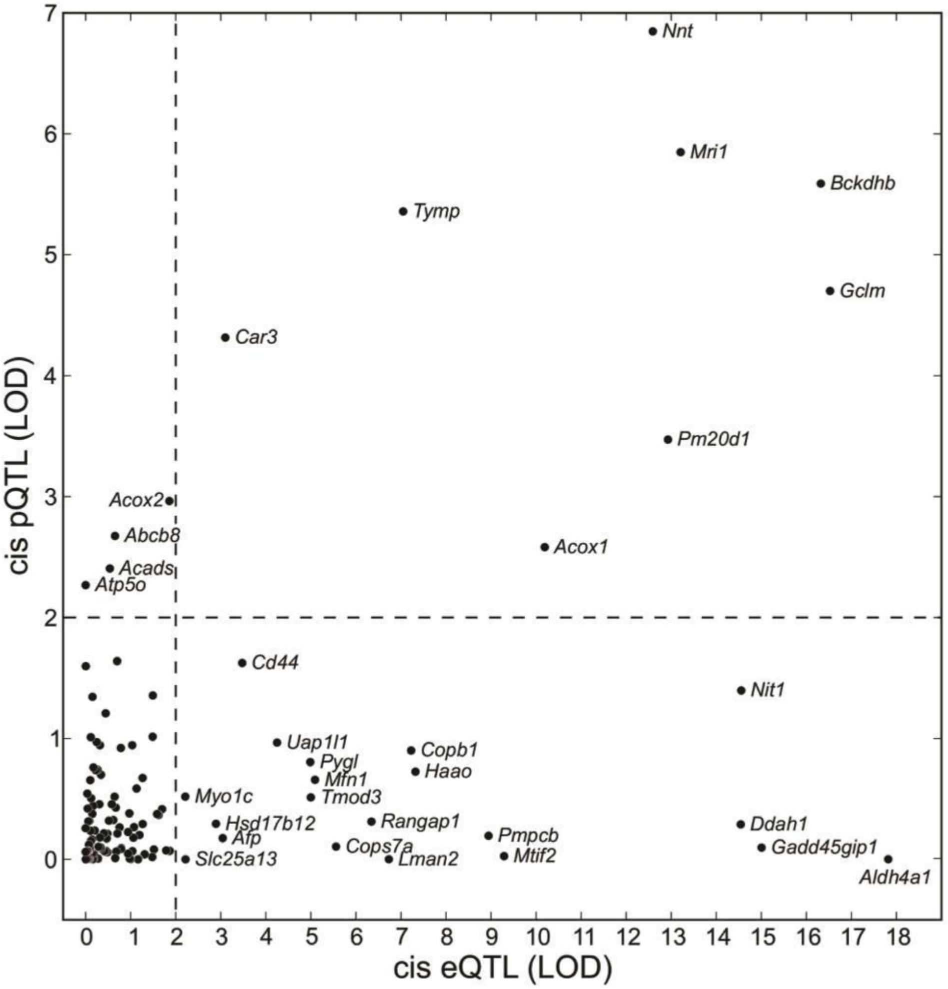


Figure 3-8. Comparison of cis-acting variation at transcript versus protein levels

Notes. The *X*-axis and *Y*-axis represent LOD scores for genes and cognate proteins measured at their closest markers (*cis* LOD). A LOD of 2 (dashed line) roughly corresponds to a nominal $p < 0.01$.

(3–4 fold difference) and a strong *cis* eQTL (chr3:145 Mb, LOD ~ 14.5) favoring the *B* allele. However, the protein difference across BXDs does not map to the location of gene and protein difference between *B* and *D* alleles has a one-tailed *p* of 0.2—a reasonably strong negative result. This case is doubly interesting because variation in DDAH1 protein maps as a *trans* pQTL (Chr7: 27.85 Mb, LOD ~ 2.5).

Majority of aberrant alleles do not affect expression severely

Nonsense mediated decay (NMD) is a molecular surveillance mechanism that selectively degrades aberrant transcripts produced as a result of nonsense or splice-site variants [234-236]. NMD of aberrant transcripts should result in extreme allelic ratios (close to zero or one). However, over two-thirds of nonsense variants (transcripts) in human cell lines escape NMD through unknown mechanisms [237]. We measured allelic ratios for 12 nonsense variants (transcripts) and remarkably only two—*Gbp11* (0.05, high *D*) and *Mug2* (0.90, high *B*)—had extreme ratios across multiple SNPs. Interestingly, half of the stop codon losses identified in the *B* allele only add one to two amino acids to the variant proteins, including VMN2R79 (+1 amino acid), ADAM3 (+1), SPNS3 (+2), ZCCHC9 (+2), DLGAP5 (+2), and HOGA1 (+1). None of these transcripts have extreme allelic ratios. We found that a third of murine genes have one or more in-frame stop codons in close vicinity (<30 nucleotides) to the original stop codon. Tandem stop codons are also known to be conserved in yeast [238], and may provide a safeguard against stop codon losses. We also evaluated 36 splice-site variants and only five of these transcripts, including *Cyp2c39* (0.02), *Arhgef10* (0.03), *Pik3c2g* (0.05), *Lox14* (~0.9), and *Rpsa* (0.99) had extreme ratios.

In conclusion a majority (> 85%) of presumed aberrant transcripts including nonsense and splice-site variants escape NMD. We speculate that the use of alternative stop codons or splice sites in the immediate vicinity of the primary mutation apparently prevents aberrant transcript production.

Mechanistic insights into the basis of allele-specific expression—quantitative and qualitative differences

Cis-acting variants affect expression in three major ways: (1) by modulating transcription rates and stability (mRNA abundance), (2) by modulating transcript processing (splicing and polyadenylation), and (3) by altering mRNA transport and storage [239]. Allelic ratios of SNPs that represent different regions of a transcript can be collectively analyzed and compared to provide mechanistic understanding of these alternative mechanisms. Multiple SNPs that have the same polarity and roughly the same magnitude of effect suggest variants in enhancers or transcription-factor binding sites that control transcript levels globally. For example, *Gclm*, a gene involved in the metabolism of dietary lipid [240], that also has a strong *cis* pQTL (LOD ~5) in liver with high expression of the *B* allele. All five SNPs have ASE with the same polarity. Another example is *Nnt*, a gene linked to insulin hypersecretion in the *D* parent [241], has a strong *cis* pQTL (LOD ~8) in liver with high expression of the *D* allele. All eight SNPs exhibit significant ASE and with the same polarity (**Figure 3-9a**).

Neighboring SNPs located within 5’ or 3’ UTRs that have opposite polarity suggest allele-specific differential usage of alternate transcriptional initiation or polyadenylation sites. One example is *Txndc9*, a gene linked with colorectal cancer in humans [242]. This gene has multiple transcripts with alternative polyadenylation sites as demonstrated by multiple mRNAs in RefSeq and Ensembl gene models. Two SNPs located in exons 1 and 3 have significantly ASE with high *B* expression whereas eight SNPs located in the extended 3’ UTR (**Figure 3-9b**) have high *D* expression suggesting allele-specific differences in 3’ UTR processing. A similar pattern is observed in array data: probe sets in coding exons have high *B* expression whereas those in the 3’ UTR have high *D* expression. The longer 3’ UTR of the *D* allele harbors putative binding sites (PhastCons > 0.5 and mirSVR score < -0.3 [243]) for multiple miRNAs, including *miR-539*, *miR-96*, *miR-129-5p*, that may explain overall low expression of the *D* transcript. Another example is *Slc38a3*, a glutamine transporter involved in ammonigenesis [244,245]. GenBank and Ensembl gene models demonstrate multiple transcripts that use alternative transcriptional initiation sites. Four SNPs in exons and a SNP in 3’ UTR have high *D* expression. However, SNPs in the 5’ UTR have variable ASE (**Figure 3-9c**). Two SNPs (*rs30029220* and *rs29646102*) exclusive to the longer 5’ UTR have high *D* expression whereas a SNP (*rs3672647*) in the shorter 5’ UTR has high *B* expression suggesting that the *D* allele favors usage of transcript with the longer 5’ UTR.

Finally, SNPs with opposite ASE polarity in different coding exons are probably caused by alternative exon usage or alternative splicing. For example, carbonic anhydrase 3 (*Car3*), a gene linked to adipogenesis [246], has a strong *cis* pQTL (LOD ~5) with high expression of the *D* allele. Five SNPs located exclusively in a long isoform show significant ASE with high *D* expression: one SNP in the exon 6 and four SNPs in the 3’ UTR (**Figure 3-9d**). In contrast, 5 SNPs located exclusively in the short isoform have high *B* expression.

DISCUSSION

Allele-specific expression differences are a major driver of phenotypic differences and variation in disease risk. We exploited RNA-seq and eQTL data sets to quantify the extent and intensity of cis-acting variation in expression in liver. After correcting for alignment bias, we achieved the expected symmetrical distribution of allelic differences. Well separated SNPs within single exons are highly concordant both in strength and polarity of effects.

Having dealt with these technical challenges, we were able to identify statistically significant ASE differences with minimum fold difference of 1.25x for nearly half of all assayed transcripts. This latter finding strongly supports recent work by Crowley and colleagues [204] demonstrating pervasive and high levels of ASE in brain and other tissues. In each F1 strain contrast, they detected significant ASE in 50% or more of all

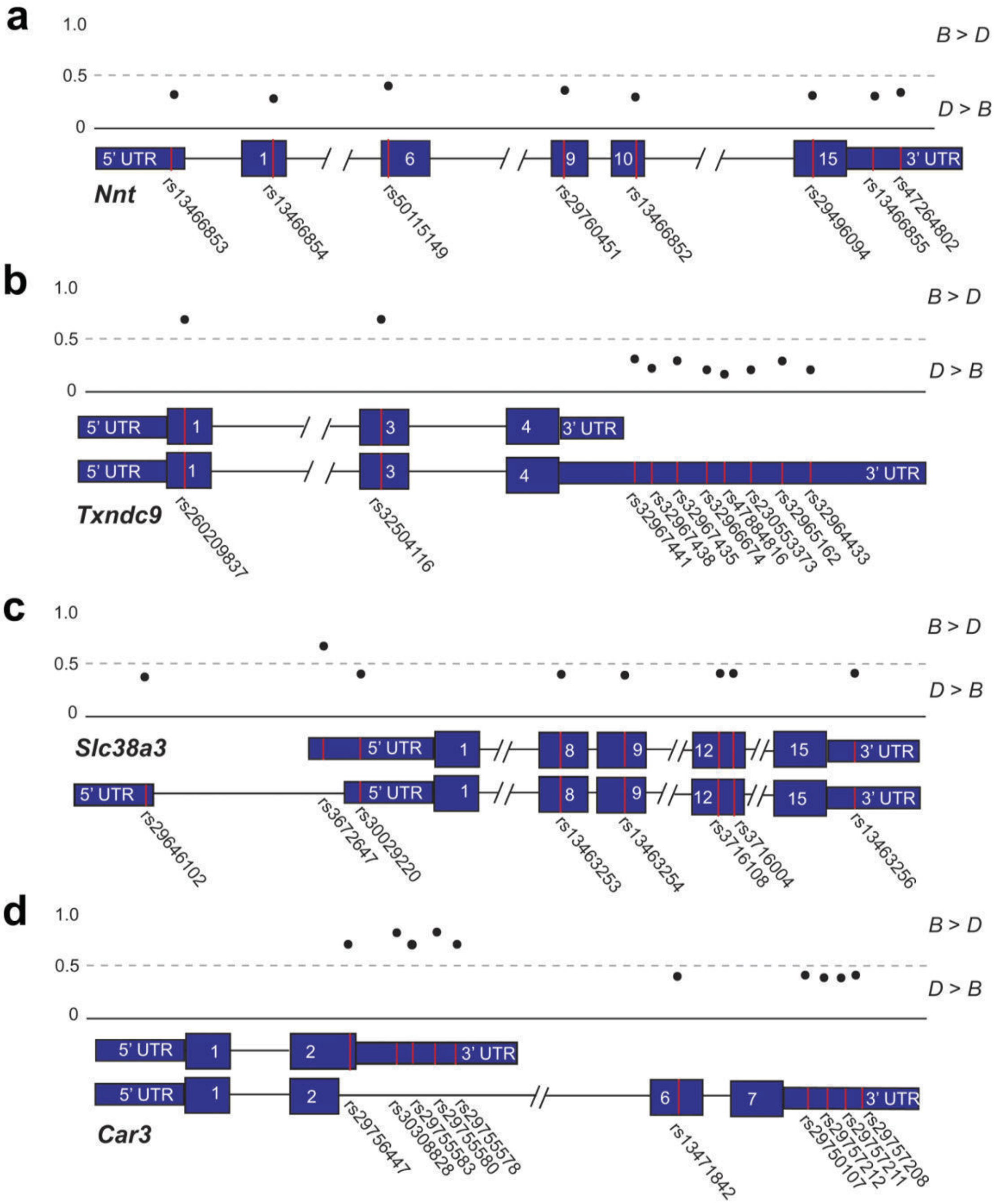


Figure 3-9. Schematic examples of genes potentially associated with different categories of *cis*-regulatory mechanisms

Notes. The *Y*-axis shows the allelic ratios of SNPs located within the gene. An allelic ratio greater than 0.5 (dashed line) represents high expression of the *B* allele. Examples of allele-specific regulation of **(a)** overall gene expression, **(b)** 3' UTR processing, **(c)** 5' UTR processing, and **(d)** isoform usage.

tested genes/transcripts at an FDR of 0.05. In total, 90% of testable genes exhibited ASE effects in at least one pair of strains. Lagarrigue and colleagues [214] detected somewhat less pervasive ASE effects (~20%) in liver of C57BL/6JxDBA/2J F1 animals, but this is most certainly a matter of lower RNA-seq read depth (statistical power), and a higher fold difference (1.5X) criterion they used to identify ASE. Of 2,256 genes, they only observed 383 genes with significant ASE. We are now able to address three questions posed in the introduction.

Highly conserved genes have low levels of ASE

Do differences in the magnitude of ASE represent differences in complexity of expression control or in evolutionary history? To answer these questions we compared a group of genes with very low and very high ASE. We found no differences in the density of cis-regulatory elements, but genes with low ASE do appear to be under more intense purifying selection. Fifty percent of the non-ASE set are house-keeping genes [230] and are likely to evolve comparatively slowly [247,248]. In contrast, genes with high ASE are likely to have higher functional redundancy as estimated indirectly by numbers of paralogs, and they are also enriched in tissue-specific functions. In our study of liver they are involved in the metabolism of lipids, fatty acids, and xenobiotics. We speculate that high gene sets with higher ASE may function in tissue-specific pathways that tend to retain both higher numbers of paralogs and be under less evolutionary constraint. The comparatively high range of variation in expression of these genes may be crucial to conferring greater physiological tolerance to noise and environmental challenges. ASE may also be one of the genetic mechanisms that underlie the canalization of phenotypes [249,250]

Poor overlap between variants affecting transcript and protein abundance

Transcript abundance has been shown to correlate only modestly with protein abundance. [219,251-254]. As expected, there is considerable disparity in allelic variation detected at mRNA and protein levels. A few of the *cis* eQTLs with very high LOD scores (≥10) but essentially no *cis* pQTLs (LOD score < 1) are *Ddah1*, *Gadd45gip1* and *Aldh4a1*. Fu and colleagues suggested an increased buffering at the level of proteins and metabolites, such that only a few genetic variants modulate major phenotypic variation and majority of them remain silent [255]. Factors that are known to contribute towards the disparity between mRNA and protein levels include post-transcriptional and post-translational modifications [254], differences in half-lives [256], variability in mRNA expression level due to changes in cell-cycle [257].

Comparison between ASE and eQTL mapping results

We found significant overlap in cis-modulated genes identified by ASE and eQTL analysis, despite substantial differences between methods and assays. Eighty percent of *cis* eQTL genes are also detected by ASE, and ~90% of them have the same polarity. The set of 683 genes identified by both methods have significantly higher LOD scores than the set of 184 genes identified only by eQTL mapping. In our work, when ASE methods fail to detect known eQTLs, this is almost certainly due to inadequate read depth (statistical power). High sampling error in RNA-seq data will affect power of ASE analysis especially for genes with low expression [82,258]. As shown above, high read depth is required to detect small allelic difference. A small fraction of presumed *cis* eQTLs can be local trans eQTL effects of neighboring genes.

The jointly identified set of ~650 genes has greater allelic differences than the set of 1,125 genes identified only by ASE. A large fraction of subtle allelic differences identified only by ASE may have been confounded by noise or epistatic trans-acting effects in the eQTL analysis. The small sample size of the BXD cohort used for the eQTL mapping may not have adequate statistical power to map weak *cis* eQTLs, especially in the presence of epistatic *trans* eQTLs. Additionally, the LOD threshold of greater than 3 used to define cis eQTLs may be too stringent in this particular context.

To the best of our knowledge, Babak and colleagues [259] were the first to compare F1-derived ASE results with eQTL results from F2 intercrosses for adipose and islets samples. They found an 80% overlap between genes exhibiting ASE and genes with *cis* eQTLs. Lagarrigue and colleagues [214] found a 60% overlap between the methods. Hasin-Brumshtein and colleagues [260] performed a similar comparison in adipose tissue and reported relatively poor overlap (~20%), but as noted above, differences with our more concordant results are most likely cumulative result of differences in criteria and ratios of statistical power of ASE analysis using F1 hybrids and cis eQTLs analysis using large intercrosses.

As highlighted in the introduction, it is now clear that most common variation in phenotype and disease risk are linked to variants that modulate patterns of gene expression. ASE is a sensitive and a cost-effective method to detect cis-acting differences in expression. Environmental and trans-acting factors are fully controlled in isogenic F1 individuals, and ASE analysis only requires a small F1 sample size. In this respect it has a clear advantage over eQTL analysis of segregating populations. However, many classical laboratory strains have been derived from ancestral stock with limited haplotype diversity. As a result, a large fraction of an F1 genome will be identical by descent (IBD) and genes in these regions cannot be interrogated using ASE.

Linkage-based eQTL analysis adds two important dimensions to an ASE study. First, it makes it possible to assign causality to specific variants using high-resolution mapping populations [88,89]. Second, eQTL analysis makes it possible to study the downstream effects of differential expression. These downstream effects are detected as *trans* eQTLs of other mRNAs or proteins.

ASE varies between different environments and genetic backgrounds

Estimates of ASE and *cis* eQTL will vary as a function of genetic background [204], tissue [172], environment [219], and sex [261]. For example, cis eQTLs effects can be strongly dependent on diet. The *cis* eQTL associated with *Ndusf2* increases from LOD score of 2 in a mouse cohort on a normal chow diet to a LOD score of 6 in a cohort on a high fat diet [219]. For these reasons, one should not expect estimates of ASE in liver of one population or treatment to generalize. Nevertheless, many of the large ASE effects caused by strong cis-acting variants will often be well conserved across environments, cell types, and genetic backgrounds. For example, ASE effects due to copy number variants [199], retrotransposons disrupting 3' UTRs [88], and nonsense mutations [262] will often produce strong and consistent ASE effects across many tissues and treatments.

CHAPTER 4. SYSTEMATIC PHENOME-WIDE ASSOCIATION USING BXD RECOMBINANT INBRED STRAINS OF MICE TO TEST GENE FUNCTION AND DISEASE RISK

SYNOPSIS

Phenome-wide association is a novel reverse genetic strategy to analyze genome-to-phenome relations in human clinical cohorts. Here we test this approach using a large murine population segregating for ~5.5 million sequence variants, and we compare our results to those extracted from a matched analysis of gene variants in a large human cohort. For the mouse cohort, we amassed a deep and broad open-access phenome consisting of 4,230 metabolic, physiological, pharmacological, and behavioral traits, and more than 80 independent eQTL transcriptome, proteome, metagenome, and metabolome datasets—by far the largest coherent phenome for any experimental cohort (www.genenetwork.org). We tested downstream effects of subsets of variants and discovered several novel associations, including a missense mutation in fumarate hydratase that controls variation in the mitochondrial unfolded protein response in both mouse and *C. elegans*, and missense mutations in *Col6a5* that underlies variation in bone mineral density in both mouse and human. Unlike genome-wide association, negative results in a deep phenome scan can be informative. Downstream effects of allelic variants with presumed deleterious effects on protein structure or mRNA and protein levels are often small or undetectable.

INTRODUCTION

Identifying sequence variants that control sets of linked phenotypes is fundamental to understanding the molecular basis of both Mendelian and complex traits [172,263-265]. A variety of reverse genetic approaches to induce loss- and gain-of-function have been used to causally tie DNA variants to discrete phenotypes [266]. However, reverse genetics presents two challenges. The first is evaluating a potentially broad spectrum of phenotypes, biomarkers, and endophenotypes that are downstream of sequence variants at different stages of development and under different conditions. The second is evaluating the impact of these variants across different genetic backgrounds that influence trait penetrance. Phenome-wide association studies (PheWAS) have been developed recently to address both challenges [59,267]. In order to establish the first murine resource for phenome-wide association we have used a large cohort of strains—the BXD family—that we generated by crossing two fully inbred strains—C57BL/6J and DBA/2J. We have discovered novel genetic associations and have translated them to a clinical cohort through human PheWAS using electronic health record (EHR) data. The use of systematic phenome scanning provides an exemplar of an effective paradigm for genome-to-phenome mapping and the analysis of pleiotropic genetic effects.

MATERIALS AND METHODS

C. elegans experiments

C. elegans were cultured at 20°C on nematode growth media agar plates seeded with bacteria. Strains were provided by the Caenorhabditis Genetics Center (University of Minnesota). The strains used were SJ4100 (zcIs13[hsp-6::GFP]), SJ4005 (zcIs4[*hsp-4*::GFP] and dvIs70 [hsp-16.2p::GFP + rol-6(su1006)]. RNAi constructs were isolated from the RNAi feeding library (GeneService) and experiments were carried out using standard feeding methods. The identity of each RNAi clone was verified by sequencing. RNAi treatment was started from embryonic stage. GFP was monitored in day 1 adults. Worms were immobilized with 6 mM solution of tetramisole hydrochloride (Sigma) in M9 and imaged using Nikon DS-L2 fluorescent microscope.

Organization and categorization of mouse phenome

The BXD Phenotype database has been amassed by literature review, direct data entry by our team, and by collaboration with many investigators. Data are reviewed prior to entry in GeneNetwork by the senior author. Phenotypes are currently split into fifteen broad phenotypic categories (www.GeneNetwork.org). Phenome curation and description was initiated by RWW and Dr. Elissa Chesler in 2002 by literature review and data extraction. The early work is described briefly in Chesler and colleagues [268,269]. We have used a controlled vocabulary and set of rules described here (http://www.genenetwork.org/faq.html#Q-22). Descriptions include a "prefix" of major biological and domain categories such as "central nervous system", "cancer biology", "immune system". These domains have been used to define major categories used in figures and graphs.

PheWAS analysis in mice

PheWAS were performed for a total of ~11,000 variants, including 10,895 missense, 61 nonsense, 196 splice site, 99 frame shift mutations, and 215 CNVs The closest marker for each variant from a set of 3,804 genetic markers—each representing a unique haplotype block—was used to represent that variant in the PheWAS. A total of 84 expression datasets were used for calculating the number of mRNA assays. Among these, 16 datasets highlighted in grey were further used for molecular phenome scan analysis. Similarly, we used 4,236 classic phenotypes from GeneNetwork.org (www.genenetwork.org) to study the association between variants and phenotypes. We calculated the *p* value of the Pearson correlation between each marker (variant) and 4,236 phenotypes and ~40,000 transcripts for the expression data. All *p* values of correlation were calculated as a two-tailed test, and the *q* values (false discovery rate; FDR) were calculated using QVALUE[270]. We used an FDR threshold of 0.01 for associations. The analyses were performed using in-house Python scripts, and the R statistical package.

PheWAS analysis using GeneNetwork.org

The PheWAS analysis can be performed in GeneNetwork (www.GeneNetwork.org) by simply following these three steps. We use an example of a missense variant (Chr2: 25,398,350) in *Entpd2*: **Step 1:** Find a marker closest to the high impact sequence variant of interest identified by our deep resequencing of the DBA/2J genome. We can search for the closest marker located on Chromosome 2 at 25.39 Mb by searching the BXD Genotype data set using a string such as "Position=(Chr2 25 26)", where 25 and 26 are the proximal and distal search regions on Chr 2 measured in megabases. This will return a list of markers—mainly SNPs and microsatellites—close to the high impact sequence variants in *Entpd2*. From the list of markers, we selected *rs8250941*, located within ~100 Kb of the high impact sequence variants. **Step 2:** Compute correlations between this marker and all ~4000 BXD phenotypes. In the case of *rs8250941*, GeneNetwork traits 10015 and 10014 are highly correlated with this marker and have –logP association scores larger than 10. Traits that have peak association scores (LRS or LOD) that are very close to the location of sequence variants are candidate traits. **Step 3:** Compute correlations between the marker and all or a subset of relevant transcriptome or proteome data sets. This may produce large output tables with as many as 20,000 endophenotypes. However, these can be sorted easily by the position of the peak association score (click on column that is labeled Max LRS Location). When correctly sorted, this table will highlight those candidate mRNAs that presumably map to sequence variants at the *Entpd2* locus.

PheWAS in humans

PheWAS for human data was performed using 29,722 individuals with Illumina HumanExome array data identified as European ancestry in the EHR and by using *Structure*[271]. To define diseases, we mapped International Classifications of Diseases, 9th edition, (ICD9) codes from the EMR into 1,645 possible PheWAS phenotypes using methods described previously[267]. PheWAS phenotypes aggregate like ICD9 codes together (e.g., type 2 diabetes codes as a specific phenotype), are hierarchical (e.g., "inflammatory bowel disease" is a parent of "Crohn's disease" and "Ulcerative colitis"), and include logic to select controls for each case definition. We considered only phenotypes with at least 20 individuals for analysis, and required each case to have at least 2 ICD9 codes for a PheWAS phenotype to be considered a case (those with only 1 code are neither a case nor a control). Each SNP-phenotype association test was run with PLINK[272] using logistic regression adjusted for age, sex, and the first three principal components as calculated by EIGENSTRAT using ancestry informative markers. Analysis was performed assuming an additive genetic model. These data were aggregated and analyzed using Perl scripts and the R statistical package. A total of 1501 phenotypes were considered, for a per-SNP Bonferroni correction of $0.05/1501=3.3\text{x}10^{-5}$.

We then performed PheWAS for missense SNPs for each of the target genes from the mouse PheWAS that had minor allele frequencies >1% and passed quality control filters of SNP call rate >95% and sample call rate >99% in unrelated samples. SNPs were found for *ENPTD2* (*rs34618694*), *COL6A5* (*rs1353613*, *rs79867908*, *rs12488457, rs113396273, rs35886424, rs1453241, rs1497312, rs11917356, rs76864445, rs16827497, rs16827168, rs819085, rs9883988, rs61744488*), *AHR* (*rs2066853*), *ALAD* (*rs1800435*), *and HDH3* (*rs1043836*). No SNPs were available for *FH1*.

RESULTS

Phenomes

Phenome data were generated using a large cohort of recombinant inbred strains—the BXD family—that was generated by crossing two fully inbred parents—C57BL/6J and DBA/2J. Members of the BXD family collectively segregate for all sequence variants that distinguish the two parents—and in this cross these are by definition common variants. There are also interesting rare but still undefined alleles unique to each family member. The level of both genetic and phenotypic variation between parents and among the strains is high (**Figure 4-1a**). This BXD phenome includes ~4,500 quantitative metabolic, physiological, pharmacological, toxicological, morphometric, and behavioral phenotypes (**Figure 4-1b**). These traits are almost all quantitative (unlike electronic health care datasets) and have been systematically grouped into 15 major phenotype categories (www.GeneNetwork.org). We have also generated and assembled a large molecular phenome that includes expression phenotypes from ~90 large open access expression quantitative trait locus QTL (eQTL) studies generated over the past decade (www.GeneNetwork.org). On average 1.5×10^6 mRNA, 1.7×10^4 proteomic, and 6.8×10^3 metabolomic assays are available per strain (**Figure 4-1b**). Most phenotypes vary markedly across strains within the family. For example, effect of high-fat and low-fat diets on adult body weight varies substantially across genotypes (**Figure 4-1c**). Similarly mRNA and protein expression of, for example, *Bckdhb* and many other mRNA, proteins, and metabolites vary greatly (**Figure 4-1d**) [219]. The online availability of well-organized classic and molecular traits from the BXD family (see www.genenetwork.org) provides the foundation for multiscalar phenome scans of any putatively functional sequence variant.

The human phenome used in this study is a large electronic health record (EHR)-linked cohort, BioVU. BioVU currently contains >190,000 DNA samples linked to de-identified medical records to provide a large, clinically-relevant human resource to study genotype-phenotype associations; 29,722 of these individuals have extant exome variant data, which was used for matched mouse-to-human PheWAS in this study. Informed consent was obtained from all human participants.

Phenome-wide association analysis in mouse

The functionally important variants (i.e. nonsense, missense, splice site, frameshift, and CNVs) were selected for subsequent Phenome-wide association study

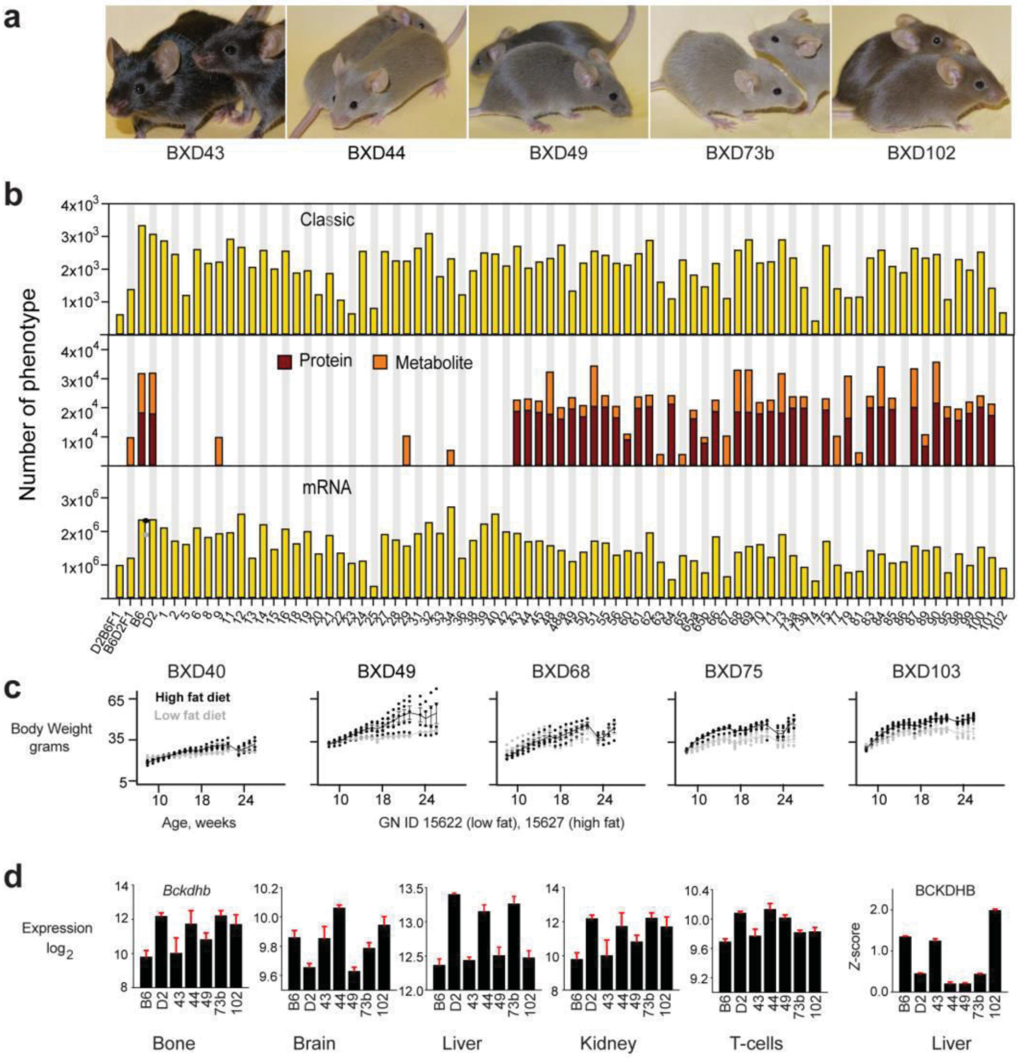


Figure 4-1. Overview of phenome data for the BXD cohort

Notes. **(a)** Five pairs of isogenic BXD cohort strains—BXD43 to BXD102. There are now approximately 100 readily available BXD strains and another 50 that are almost fully inbred. Almost all current phenome data is restricted to the parents, F1 hybrids, and BXD1 through BXD102. **(b)** Phenome data categorized by type, including classic phenotypes (top), metabolic and proteomic trait data (middle), and independent mRNA expression assays (bottom, n = 86 unique eQTL data sets). **(c)** Body weight data for BXD strains on high fat (gray) and low fat (black) diets. **(d)** Expression of *Bckdhb* mRNA and its protein in six tissues for the five BXD strains.

(PheWAS) analysis (**Figure 4-2**). We used 3,805 genotypes that represent distinct haplotype blocks in the BXD family to perform PheWAS against 4,230 classic traits as well as 602,746 endophenotypic traits from 16 distinct tissues. This analysis yielded ~14 million genotype-to-phenome correlations and ~2.0 billion genotype-to-endophenotype correlations. A total of 95 genotypes are significantly associated with 321 phenotypes, corresponding to 108 phenotypic groups, at a stringent *q* value threshold of <.01 (**Supplementary Table 4-1**). In addition, we performed differential expression analyses between the C57BL/6J and DBA/2J strains for each association by using transcripts from 16 tissues and proteins from hippocampus.

We interrogated the associations for 11,466 functionally important variants, including 10,895 missense, 61 nonsense, 196 splice site, and 99 frame shift mutations, and 215 CNVs, by mapping these variants to the nearest genotype markers within ±1 Mb. We found 650 functionally important variants associated with 97 classic phenotypes, including 634 missense variants associated with 62 phenotypes.

Examples of variant-phenome association

Among 321 classic phenotypic associations meeting a stringent *q* value threshold of < 0.01 (**Supplementary Table 4-1**), a few variants, such as those in *Gpnmb, Comt* and *Ahr,* have been associated previously with disease[88,146,273] using traditional forward genetics approaches, but the vast majority of variants have not been previously linked to any phenotype. Here, we provided four PheWAS examples, including three missense variants (*Fh1*, *Col6a5*, *Entpd2*), a nonsense variant *(Ahr)*, and a CNV *(*a region covering both *Alad* and *Hdh3)*.

The first example is a missense variant (A296T) in the fumarate hydratase mitochondrial enzyme located on chromosome 1 at 175.60 Mb (*Fh1*; **Figure 4-3a**). *Fh1* catalyzes the hydration of fumarate to malate in the tricarboxylic acid (TCA) cycle and has been linked to renal cell cancer[274]. The missense variant in the lyase 1 domain is associated with a ~1.4-fold effect on expression of *Fh1* across many tissues, including midbrain, hypothalamus, striatum, and spleen (**Figure 4-3b**). This variant is strongly associated with *Fh1* mRNA expression, as well as the expression of other mitochondrial genes, including *Mrpl50*, *Sirt3* and *Dlst* (**Figure 4-3c**). Expression PheWAS shows that the *Fh1* locus modulates mRNA expression levels of 113 mitochondrial proteins, in addition to 8 genes linked to renal necrosis, and 7 genes involved in mTOR signaling, consistent with the known role of *FH1* in renal cancer. Interestingly, four mitochondrial genes, *Hspd1*, *Hspa9*, *ClpX*, and *Lonp1* that all encode components of the mitochondrial unfolded protein response (UPR^{mt})[87]—a still poorly characterized mitochondrial stress response pathway in mammals—show strong association with *Fh1* (**Figure 4-3d**). There is furthermore a significant correlation between *Fh1* transcript levels and principal component scores of a group of UPR^{mt} genes in mouse (**Figure 4-3e**). In contrast, no genes involved in the cytoplasmic heat shock response (HSR) or the ER unfolded protein response (UPR^{er}) are associated with *Fh1*, indicating a selective association between *Fh1* and UPR^{mt} in mammals (**Figure 4-3d**). To validate this association, we examined the

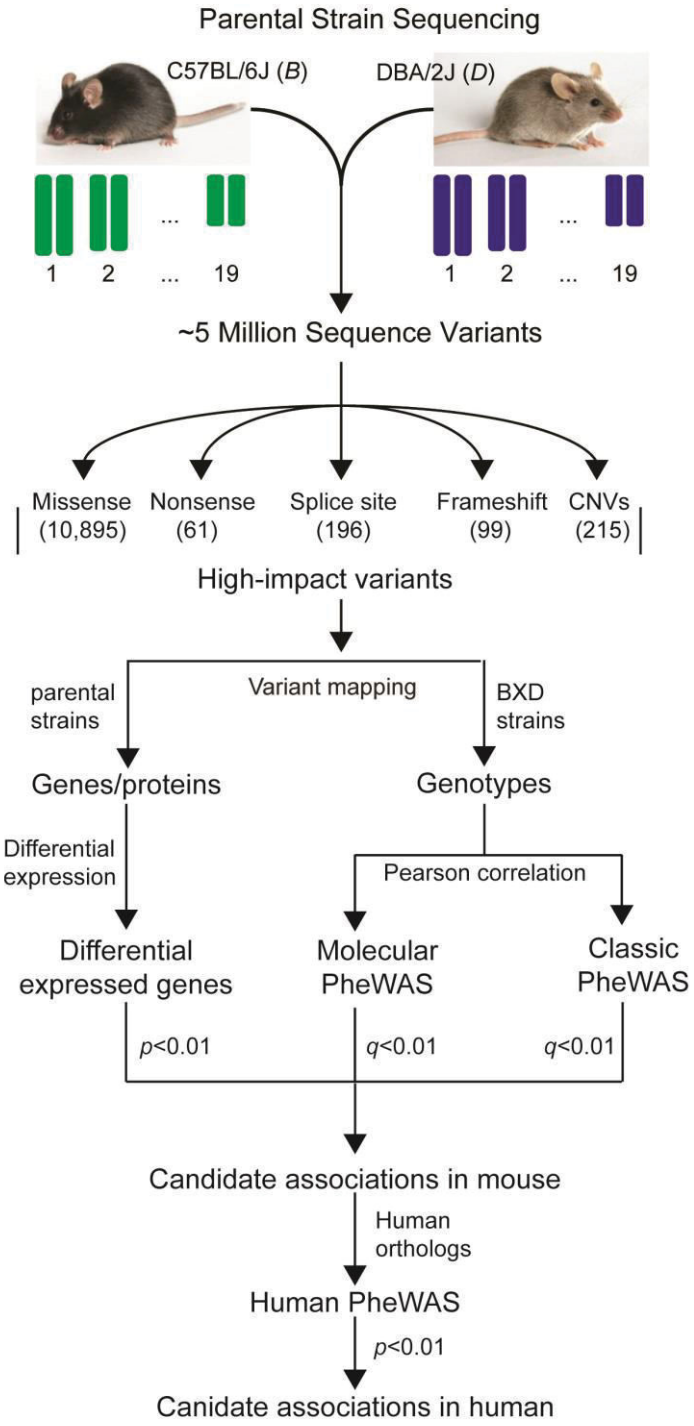
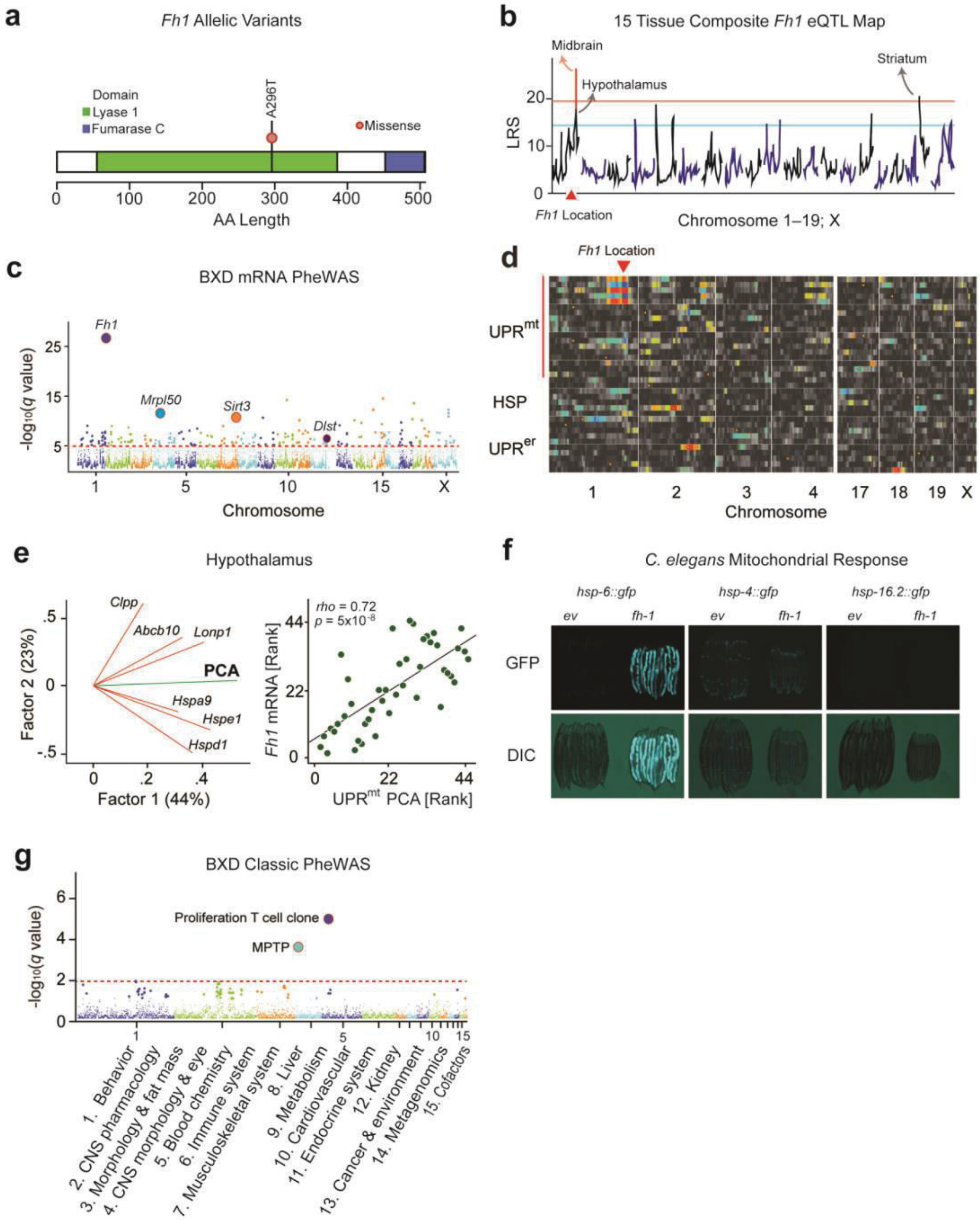


Figure 4-2. Experimental PheWAS analysis using the BXD cohort

Notes. The list of strong sequence and structural variants were selected for phenome scanning. The Pearson product-moment correlation was used to calculate the association between variants and classic and molecular phenotypes from 16 tissues. FDRs were calculated based on *p* values using QValue method. We used a *q* threshold of < 0.01.

Figure 4-3. Association analysis for a missense variant in *Fh1*

Notes. **(a)** Structure of the *Fh1* gene showing lysate 1 and fumarase c domains, the latter of which contains a missense mutation. **(b)** Combined eQTL mapping of *Fh1* mRNA across 15 tissues. The eigenvalues associated with the first principal component map to *Fh1* with a likelihood ratio statistic (LRS) of >20. The solid red line represents genome-wide significance. The yellow triangle indicates the genomic position of *Fh1*. Three peaks associated more specifically with single tissue types are labeled. **(c)** Manhattan plot of a expression phenome scan of molecular traits linked to the *Fh1* locus in midbrain. In addition to *Fh1* transcript, mRNAs for three mitochondrial-related genes (*Mrpl50*, *Sirt3*, and *Dlst*) are significantly associated with *rs4136041*. The y-axis shows the –log10 q values of association, and the x-axis shows positions of 55,681 probe sets from Agilent SurePrint array. **(d)** QTL heat map of mRNAs involved in the unfolded protein response. The x-axis lists mouse chromosome numbers, compressed for illustrative purposes. Each horizontal line represents the QTL map for a single transcript in midbrain. Transcripts are grouped into three major categories—genes involved in the canonical UPRmt, and those encoding chaperones and heat shock proteins (HSP) and those involved in unfolded protein responses in the endoplasmic reticulum (UPRer). The subset of UPRmt genes at the top are strongly modulated by the *Fh1* locus on Chr 1 (the intense colors to the upper left). In contrast, none of the UPRer subsets are modulated by *Fh1*. **(e)** Principal component analysis (PCA) plot for six UPRmt transcripts (left panel). The first two components explain ~67% of the variance in *Fh1* expression in hypothalamus. There is significant correlation between UPRmt expression and *Fh1* ($p = 5 \times 10^{-8}$) (right panel). **(f)** Validation that fumarase hydratase selectively controls the UPRmt in C. elegans. The left-most pair of images demonstrates effects of the fum-1 RNAi knockdown on hsp-6::gfp signal—a marker of UPRmt induction. Top images are GFP fluorescence; bottom panels are matched differential interference contrast (DIC) image with GFP. The middle and right panels demonstrate that the fum-1 knockdown does not induce either the UPRer (hsp-4::gfp), or the cytoplasmic heat shock response (hsp-16.2::gfp). **(g)** Manhattan plot of a phenome scan. The phenome has been subdivided into 15 categories based on biological function or tissue type. The y-axis shows the –log10 q values of ~4,230 phenotypes, and the x-axis shows the 15 categories.


a
Fh1 Allelic Variants
Domain
Lyase 1
Fumarase C
A296T
Missense
0 100 200 300 400 500
AA Length
b
15 Tissue Composite Fh1 eQTL Map
Midbrain
Hypothalamus
Striatum
LRS
0 10 20
Fh1 Location
Chromosome 1–19; X
c
BXD mRNA PheWAS
Fh1
Mrpl50
Sirt3
Dlst
-log10(q value)
5 15 25
1 5 10 15 X
Chromosome
d
Fh1 Location
UPRmt
HSP
UPRer
1 2 3 4 17 18 19 X
Chromosome
e
Hypothalamus
Clpp
Abcb10
Lonp1
PCA
Hspa9
Hspe1
Hspd1
Factor 2 (23%)
.5 0 -.5
0 .2 .4
Factor 1 (44%)
rho = 0.72
p = 5x10-8
Fh1 mRNA [Rank]
0 22 44
UPRmt PCA [Rank]
f
C. elegans Mitochondrial Response
hsp-6::gfp
hsp-4::gfp
hsp-16.2::gfp
ev fh-1
GFP
DIC
g
BXD Classic PheWAS
Proliferation T cell clone
MPTP
-log10(q value)
0 2 4 6
1 5 10 15
1. Behavior
2. CNS pharmacology
3. Morphology & fat mass
4. CNS morphology & eye
5. Blood chemistry
6. Immune system
7. Musculoskeletal system
8. Liver
9. Metabolism
10. Cardiovascular
11. Endocrine system
12. Kidney
13. Cancer & environment
14. Metagenomics
15. Cofactors

phenotypic impact of the highly conserved *C. elegans Fh1* ortholog, *fum-1* (86% sequence similarity) on unfolded protein responses. RNAi against *fum-1* causes robust activation of the mitochondrial chaperone *hsp-6::gfp* reporter, indicative of the activation of the UPR^{mt} (**Figure 4-3f**). The response was organelle-specific, and *fum-1* RNAi does not induce either *hsp-4::gfp* or *hsp-16.2::gfp*, reporters for related to the UPR^{er} or heat shock response, respectively (**Figure 4-3f**). Thus, in the BXD family, a decrease of fumarate hydratase leads to a specific mitochondrial phenotype, characterized by an UPR^{mt}.

Fh1 is also associated with two candidate phenotypes: (1) T cell proliferation (GN ID 10237; q = 2.6x10^{-5}), linked previously to mitochondrial function[275], and (2) dopamine metabolism after treatment with the mitochondrial toxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP, GeneNetwork identifier (GNID) 15151; q =0.005). Both traits are linked to *Fh1* along with the control of mitochondrial components of a UPR^{mt} pathway (**Figure 4-3g**). No extant human genotype data are yet available for *FH*—the ortholog of *Fh1*.

The second example consists of a set of tightly linked missense variants in collagen 6A5 on chromosome 9 at 105.76 Mb (*Col6a5,* **Figure 4-4a**). *Col6a5* is a variant-rich gene and contains 21 missense variants, including a radical substitution (R1778C). Quantitative RT-PCR shows higher expression of the *D* allele than the *B* allele in bone marrow (**Figure 4-4b**). As expected expression differences are strongly associated with *Col6a5* locus in bone expression PheWAS (q = 3.5x10^{-4}, **Figure 4-4c**). Unlike *Fh1*, the high density of linked variants in *Col6a5* means that we cannot resolve effects of single SNPs, but the scan does define effects of *B* and *D* haplotypes for *Col6a5* across the phenome. We find that this polymorphic gene is associated with differences in bone mineral density (BMD; GN ID 16532; q =0.037) (**Figure 4-4d**) and quantitative micro-CT analysis confirms a marked difference in cortical BMD at the femoral midshaft between C57BL/6J (1069.6 ± 51.4 mgHA/cm^3) and DBA/2J parents (1170.8 ± 39.8 mgHA/cm^3) (**Figure 4-4e, Figure 4-4f**). In humans, mutations in collagen VI are associated with a variety of musculoskeletal abnormalities[276]. We performed a matched PheWAS in human using the BioVU resource and linked *rs113396273* in exon 3 of *COL6A5* (M56T) with osteopenia and other bone and cartilage disorders (p = 1.4x10^{-3}; **Figure 4-4g**). Like *rs113396273*, the other SNPs tested in *COL6A5* demonstrated similar patterns of associations including respiratory abnormalities and giant cell arteritis.

The third example involves missense variants—R149Q and A297T—in ecto-nucleoside triphosphate diphosphohydrolase 2 (*Entpd2*; **Figure 4-5a**). These variants in the triphosphatase domain are linked to differences in mRNA (fold-difference = 0.6, p value <.03) and protein levels (fold-difference = 1.7, p <.01) (**Figure 4-5b**) and generate strong cis eQTLs in multiple tissues (e.g., lung, q = 4.9×10^{-9}). A phenome scan highlights two enzymatic phenotypes: (1) Ca2+- and (2) Mg2+-stimulated ecto-ATPase activity (GN ID 10014 and 10015; q = 9.97×10^{-5} and .007) (**Figure 4-5c**). Both are direct measures of ATPase activity—prima facie evidence that one or both of these SNPs are causal. In the BioVU human clinical cohort, *rs34618694* in ENTPD2, is associated with microophthalmia (p = 2.4x10^{-4}) and visual defects (p = 2.2x10^{-3}; **Figure 4-5d)**.

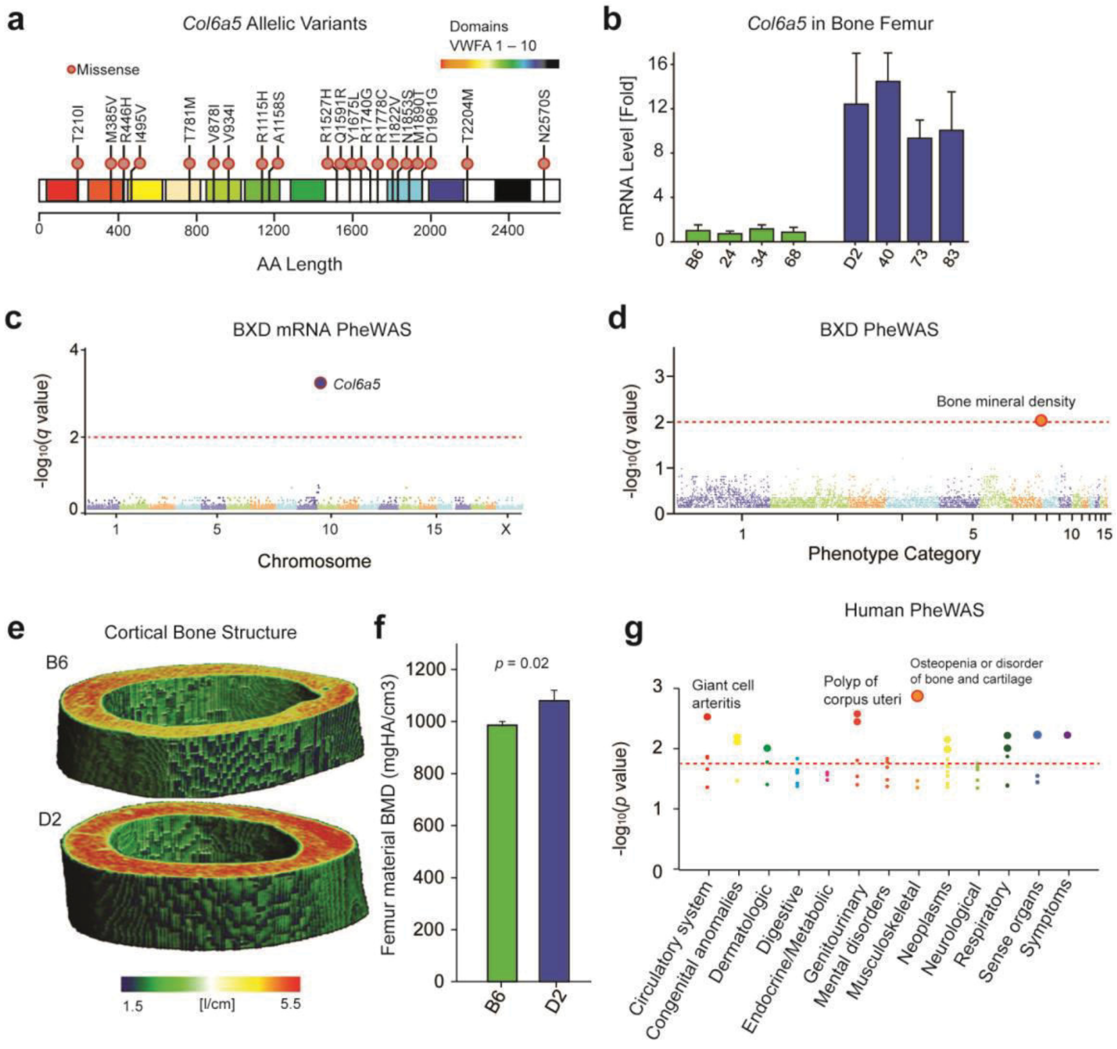


Figure 4-4. Association analysis for missense variants in *Col6a5*

Notes. **(a)** Twenty missense variants in *Col6a5* distributed across 10 von Willebrand factor A-type (vWFA) domains. **(b)** Differential mRNA expression of Col6a5 in tibias (n = 4) measured by rtPCR. The D haplotype (blue, right) has far higher expression than the B haplotype (green) relative to *Gapdh*. **(c)** Phenome scan of *Col6a5* (*rs13480398*) across mRNA assays for femur. **(d)** Phenome scan of *Col6a5* (*rs13480398*) across classic phenotypes. **(e)** Marked difference in bone density between C57BL/6J and DBA/2J parents. Femurs from 12-week-old mice were scanned using high-resolution micro-CT (μCT40, SCANCO Medical, Bassersdorf, Switzerland). More highly mineralized areas are indicated in red. **(f)** Difference in material bone density (p = 0.02; two-tailed Student’s t-test, n = 3). (g) Human phenome scan for association of *Col6a5* (*rs113396273*) across BioVU.

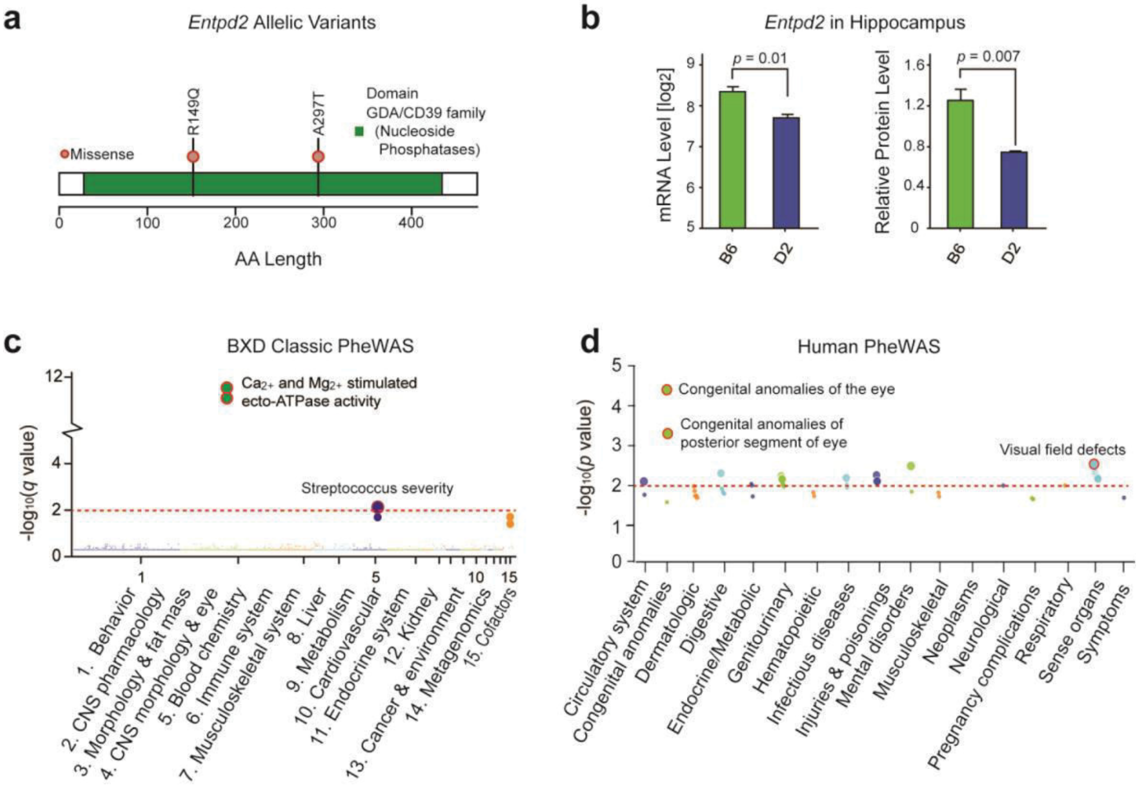


Figure 4-5. Association analysis for missense variants in *Enptd2*

Notes. **(a)** Structure of the *Entpd2* gene showing two missense mutations and GDA/CD39 family domain. **(b)** *Entpd2* expression differs at both the mRNA and protein levels between C57BL/6J and DBA/2J strains (n = 3 replicates/genotype). **(c)** Phenome scan of *Entpd2*. **(d)** Human phenome scan of *ENTPD2* across thousands of clinical records using BioVU Denny et al. (2013). The SNP *rs34618694* was selected for this scan. Three clinical traits related to eye diseases are highlighted. The red dotted line is at $p < 0.01$.

The fourth example is a high impact nonsense variant—a lost stop codon in the *D* allele of *Ahr*. *Ahr* is an important transcription factor that modulates P450 gene expression in response to xenobiotics such as dioxin[277]. Although the effects of this SNP on protein length are already known[190] (**Figure 4-6a**), the pleiotropic consequences of this mutation have not been evaluated. This variant is significantly associated with mRNA ($q = 1.7 \times 10^{-3}$; **Figure 4-6b**) and protein abundance of *Ahr* in liver ($q = 0.0085$; **Figure 4-6c**). Classic PheWAS linked this variant to the frequency with which cleft palates is induced by 2,3,7,8-tetrachlorodibenzofuran injection (GN ID 10714; $q = 3.2 \times 10^{-3}$) (**Figure 4-6d**). *Ahr* variants have also been definitively linked to differences in locomotor activity[262]. Consistent with the results of the BXD PheWAS, a matched PheWAS in humans using BioVU links *rs2066853* in *AHR* with cleft palate ($p = 0.012$; **Figure 4-6e**).

In the final example, we tested the effect of copy number variants on gene expression and phenotypes. A CNV region on chromosome 4:62.49-62.52 Mb that spans both *Alad* and *Hdhd3—*is interesting and involves a 4X expansion in strains with the *D* haplotype. The 30 kb CNV is otherwise identical by descent (**Figure 4-7a**). This CNV is linked with high variation in mRNA expression of *Alad* and *Hdhd3* in multiple brain regions (**Figure 4-7b, Figure 4-7c**), lung ($q = 2.1 \times 10^{-7}$), eye ($q = 1.3 \times 10^{-10}$), and liver ($q = 9.2 \times 10^{-4}$). Quantitative proteomics of hippocampus confirms significant upregulation (ALAD 2.3-fold, $p < 0.01$, HDHD3 1.5-fold, $p < 0.01$). The CNV expansion of *Alad* and *Hdhd3* is strongly linked to two classic phenotypes: pain response (GN ID 11307; $q = 7.8 \times 10^{-3}$) and deoxycorticosterone levels in cerebral cortex (GN ID 12568; $q = 2.6 \times 10^{-4}$) (**Figure 4-7d**). A matched phenome scan in human demonstrates that *rs1800435* in *ALAD* is associated with chronic pain ($p = 2.2 \times 10^{-2}$) (**Figure 4-7e**).

Phenotypic resilience

One surprising finding is that a large proportion of genes with variants that we initially believe would have high phenotypic impact failed to associate with any classic phenotypes, or even with molecular endophenotypes. Among 41 confirmed nonsense variants with high predicted impacts, 12 nonsense variants failed to associate with any endophenotypes (across scans of 16 transcriptome data sets in different tissues) or with classic phenotypes at $q < 0.01$. Failure to detect associated phenotypes could be interpreted as false negative results or inadequate phenome coverage, but we suspected that most commonly this reflects molecular resilience that buffers the phenotype from apparently strong homozygous mutations.

DISCUSSION

We have evaluated the phenotypic effects of a spectrum of genetic variants in a large mouse cohort by phenome-wide association. The variety and depth of phenotype data that we have assembled over the last decade for the BXD cohort make this the largest coherent multiscalar data set for any segregating population. Of course, there are

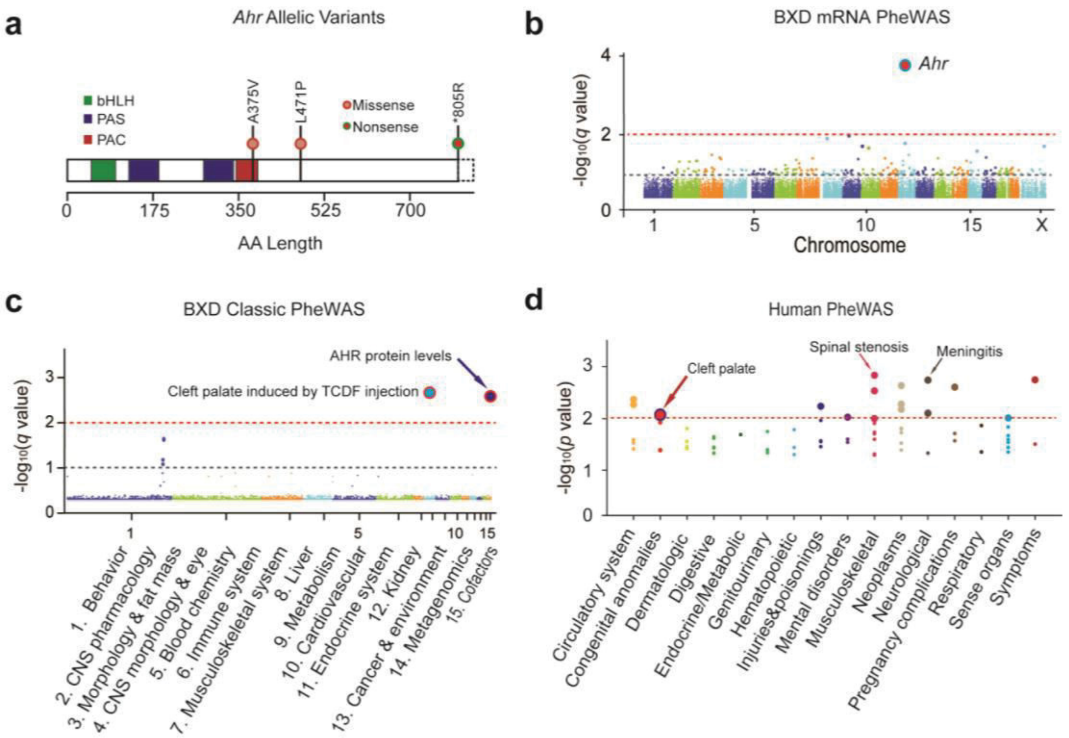
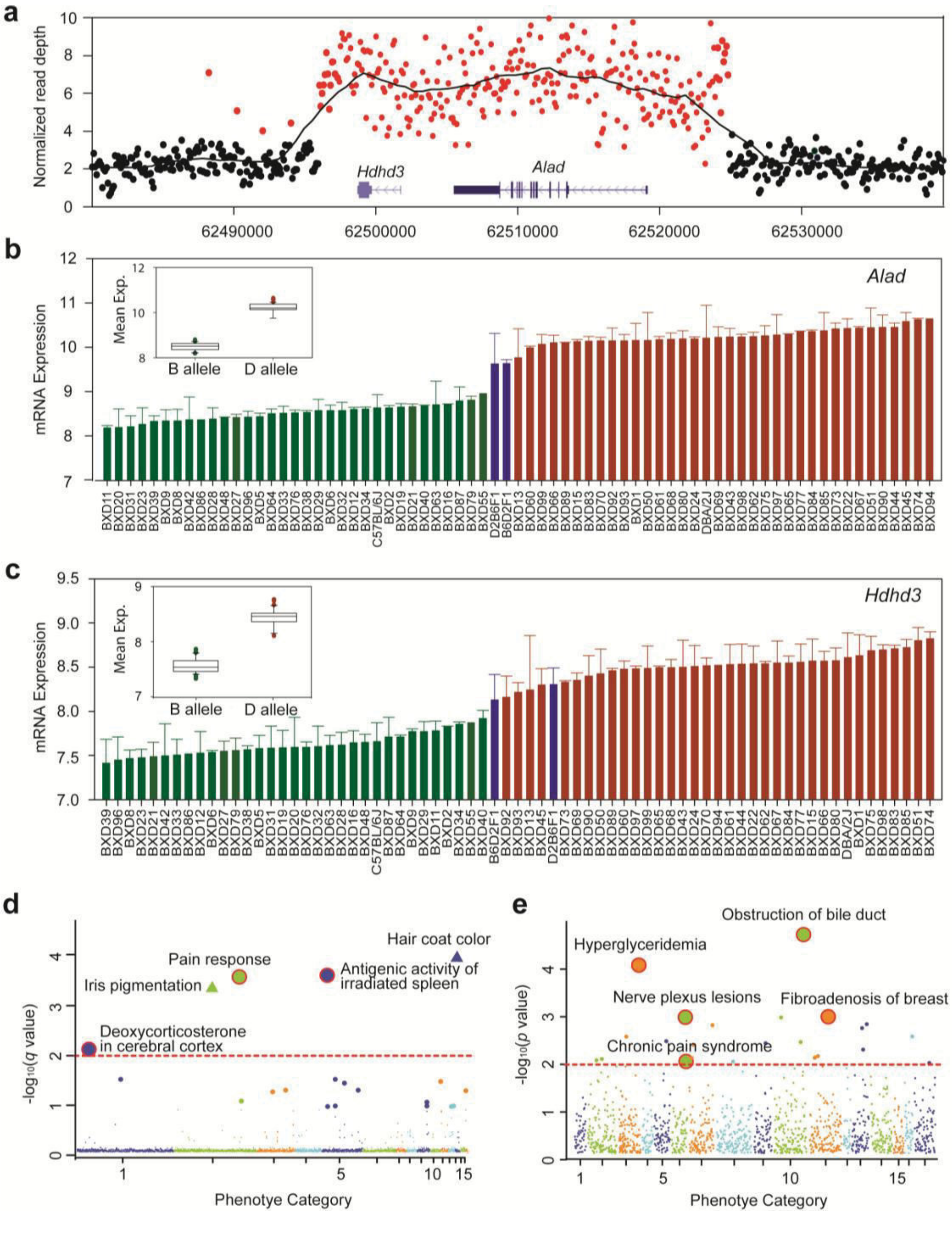


Figure 4-6. Association analysis for nonsense variant in *Ahr*

Notes. **(a)** Structure of the *Ahr* gene showing three domains with a nonsense mutation and five missense mutations. The nonsense mutation (*805R) leads to loss of the stop codon, and the addition of 43 C-terminal amino acids. Dotted rectangle to the right is the extended coding region in the D haplotype. **(b)** Phenome scan of *Ahr* (*rs3711448*) across mRNA assays for liver. **(c)** Phenome scan of *Ahr* across classic phenotypes. Both AHR protein level and cleft palate induced by TCDF injection are strongly linked to *Ahr*. **(d)** Manhattan plot showing the association in human between SNP (*rs2066853*) in *AHR* and classic phenotypes. The cleft palate phenotype is also associate with *AHR* in human clinical cohorts.

Figure 4-7. Association analysis for CNV covering *Alad* and *Hdhd3*

Notes. (a) The CNV region for *Alad* and *Hdhd3* derived using read-depth information from genome sequencing. Red dots represent at least a two-fold increase in coverage compared to the reference genome. The x-axis shows the reference genomic position of the CNV. Two gene models (i.e. *Hdhd3* and *Alad*) are shown in the CNV plot. (b and c) Rank ordered mean expression levels of *Hdhd3* and *Alad* across 67 BXD strains, their parental strains, and F1 crosses. Expression values are normalized on a log2 scale (mean ± SE). Strains with D alleles (red) have higher levels of *Alad* and *Hdhd3* compared to B alleles (green). F1 hybrids (blue) are intermediate. The comparison between B and D alleles for *Alad* and *Hdhd3* are shown in an inset boxplot. (d) The phenome scan of the BXD cohort highlights several interesting potential phenotypes including thermal nociception, brain deoxycorticosterone levels, and antigenic activity in the spleen. Two triangles represent pigmentation traits that we know they are associated with a variant in the linkage disequilibrium block. (e) Manhattan plot showing the association in human between a SNP (*rs1800435*) in ALAD and classic phenotypes.


a
Normalized read depth
10
8
6
4
2
0
Hdhd3
Alad
62490000
62500000
62510000
62520000
62530000
b
mRNA Expression
12
11
10
9
8
7
Alad
Mean Exp.
12
10
8
B allele
D allele
BXD11 BXD20 BXD31 BXD23 BXD39 BXD9 BXD8 BXD42 BXD86 BXD28 BXD48 BXD27 BXD96 BXD5 BXD64 BXD33 BXD76 BXD38 BXD29 BXD6 BXD32 BXD12 BXD34 C57BL/6J BXD2 BXD19 BXD21 BXD40 BXD63 BXD16 BXD87 BXD79 BXD55 D2B6F1 B6D2F1 BXD13 BXD60 BXD99 BXD66 BXD89 BXD15 BXD83 BXD70 BXD92 BXD93 BXD1 BXD50 BXD61 BXD68 BXD80 BXD24 DBA/2J BXD69 BXD43 BXD98 BXD62 BXD75 BXD97 BXD65 BXD77 BXD84 BXD85 BXD73 BXD22 BXD67 BXD51 BXD90 BXD44 BXD45 BXD74 BXD94
c
mRNA Expression
9.5
9.0
8.5
8.0
7.5
7.0
Hdhd3
Mean Exp.
9
8
7
B allele
D allele
BXD39 BXD96 BXD8 BXD23 BXD21 BXD42 BXD33 BXD86 BXD12 BXD6 BXD27 BXD79 BXD38 BXD5 BXD31 BXD19 BXD20 BXD76 BXD32 BXD63 BXD28 BXD16 BXD48 C57BL/6J BXD87 BXD64 BXD9 BXD29 BXD11 BXD2 BXD34 BXD55 BXD40 B6D2F1 BXD92 BXD93 BXD13 BXD45 D2B6F1 BXD73 BXD69 BXD90 BXD50 BXD89 BXD60 BXD97 BXD99 BXD65 BXD68 BXD43 BXD24 BXD70 BXD94 BXD61 BXD44 BXD22 BXD62 BXD67 BXD84 BXD77 BXD15 BXD66 BXD80 DBA/2J BXD1 BXD75 BXD98 BXD83 BXD85 BXD51 BXD74
d
-log10(q value)
4
3
2
1
0
Hair coat color
Pain response
Antigenic activity of irradiated spleen
Iris pigmentation
Deoxycorticosterone in cerebral cortex
1
5
10
15
Phenotye Category
e
-log10(p value)
4
3
2
1
0
Obstruction of bile duct
Hyperglyceridemia
Nerve plexus lesions
Fibroadenosis of breast
Chronic pain syndrome
1
5
10
15
Phenotye Category

an almost unlimited numbers of ways to extend this BXD phenome—from much more extensive GXE studies to single cell omics, but at the current size, the phenome is certainly large enough to explore the utility of PheWAS in an experimental population. We demonstrate that phenome scans can be effective at linking sequence variants to a range of phenotypes and can be used to identify novel and unexpected genome-to-phenome relations, or to validate hypothesized associations from independent studies. Coupling mouse and human PheWAS cohorts also shows great promise, and provides an efficient method to validate and translate key genome-to-phenome relations.

The novel associations demonstrated in this study provide insight into the genetic basis of complex traits and variation in disease susceptibility. The missense variant in *Fh1* is a case in point. A variant in *Fh1* is linked to the UPR^{mt}, a protective stress pathway specific to mitochondria, and we confirmed that downregulation of *fum-1*, the *C. elegans* homolog of *Fh1*, activates the UPR^{mt}. Various disturbances have been shown to induce the UPR^{mt}, including treatment with paraquat, a pesticide that strongly induces reactive oxygen species[278], activation of mitochondrial biogenesis[87], overexpression of aggregation-prone mitochondrial proteins[279], and interference with electron transport chain protein expression and assembly[87,280]. Here, we show that a purely metabolic perturbation, such as induced by loss of function of the TCA cycle component, fumarate hydratase, can activate the UPR^{mt}. While we have detected a single missense variant in *Fh1*, the molecular cascade that links *Fh1* to other tricarboxylic acid cycle (TCA cycle) genes (e.g. *Dlst*, *Sdha*, *Sdhb*) and a UPR^{mt} proteostasis regulatory loop requires additional analysis.

Advantages and disadvantages of PheWAS

Recent work has demonstrated that phenome scans are a powerful way to link from sequence variants to sets of phenotypes in clinical cohorts [59,267]. Here we have extended this approach to a murine cohort for which we have been generating cellular and molecular traits from many tissue and cell types and for which we can generate data on gene-by-environment interactions [75,86,219]. Despite strong functional effects of variants in humans, the minor allele frequencies are often too low to attain sufficient sample size. Murine populations such as the BXDs, the Collaborative Cross, and heterogeneous stock typically have linkage disequilibrium that is at least an order of magnitude larger than in humans. Consequently, the assignment of specific causality may be erroneous. For example, in the BXD family about 20,000 protein coding genes and 12,000 coding variants are distributed across ~4,000 haplotype blocks. Increasing the size and genetic diversity of a reference population and the number of recombination events can improve genetic resolution, but a more effective and meaningful solution, exemplified in this study, is to exploit other mouse cohorts and human cohorts for validation and cross-species translation. For example, by having multiple phenomes for a single species, along with matched databases of segregating sequence variants, it would become practical to rapidly test the replicability of genome-phenome relations. It may soon be practical to compare the BXD phenome with that of the Collaborative Cross and other large families of RI strains. Any cohort will only segregate for a subset of possible sequence variants, and variants will often not be shared across populations or species. For this reason, conservation of gene function will be a more useful currency of exchange [281,282].

While PheWAS has great potential, this approach faces several hurdles to more widespread application. The first is simply the technical and logistical challenge of generating a phenome. Intense collaborative efforts are a sine qua non even for the most tractable model organisms such as Drosophila [283]. The second is yet another example of the multiple testing problem: what is the appropriate correction given the size of the phenome and its structure? We have computed FDR *q* values at a conservative threshold and have aligned our results, when possible, with the BioVU clinical cohort. However, in both species, the selection of appropriate q values will depend on the purpose of studies and the relative costs of Type 1 and Type 2 errors. Effective solutions may require adjusting thresholds based on the scope and intent of studies, as well as prior information about gene-to-phenotype relations. Alignment of phenotype associations across both humans and mice, however, adds validity to both. Very large, densely genotyped or sequenced populations will be needed to more deeply interrogate the human phenome. The third problem is linkage disequilibrium. The intervals in which sequence variants are located is a critical factor in mapping its phenotype spectrum. Pleiotropy will be inflated as a function of gene density, SNP density, and haplotype block structure. Deconvolving contributions of linked polymorphisms will, in most cases, still require independent experimental validation and, when possible, PheWAS of human cohorts.

Phenome resilience

We searched for molecular or functional consequences of ~12,000 coding variants segregating in the BXD family, and we were surprised that only a small fraction had strong effects on mRNA and protein expression, let alone on classic phenotypes. Initially, this observation was surprising to us. A first factor contributing to phenotypic resilience comes from the depth of the phenome – in fact some phenotypic differences will only be observed under certain environmental conditions. For example, functional effects of the well-known mutation in the *Nnt* gene will be much more pronounced under metabolic stress [219]. Phenotypic resilience can also be caused by the presence of in frame stop codons or splice acceptor or donor sites in the vicinity of the original disrupted site, which may generate almost normal and non-aberrant transcripts or proteins. Additionally, some of the negative results were due to incorrect gene models that generated spurious stop codons, but even after stringently filtering both sequence data and gene models, it is clear that many strong sequence variants are successfully buffered at intermediate levels[284-286]. These silent and negative results are highly useful in evaluating the reported impact of major alleles. An example is a splice site mutation in *Cyp2c39* that inactivates this P450 enzyme and essentially eliminates expression. This mutation has no detectable impact on higher order phenotypes, a compelling negative result. An obvious explanation in this case is functional overlap with other members of the Cyp2c cluster, but we still do not have sufficient knowledge to understand the molecular basis of resilience. Another explanation is that combinations of deleterious variants in molecular and developmental network will have much more notable phenotypic effects than isolated mutations. In retrospect, this buffering of genetic variation is not surprising. A large fraction of knockout mutations in mice and other well-typed species are viable and many of these do not have any known functional consequences [287,288].

CHAPTER 5. SUMMARY

In this dissertation, I have described and performed first phenome-wide association analysis (PheWAS) on a genetically diverse murine population. PheWAS is a novel reverse genetics approach to link sequence variants to a spectrum of phenotypes and diseases, and several studies have successfully used it on human clinical cohorts [59,267]. In order to extend the approach to an experimental cohort, I used a large population of BXD recombinant inbred strains, for which we generated, assembled, and annotated key genetic, genomics, proteomics, and phenome data.

In chapter 2, I explored the genetic variation between the parental strains of the BXD population by analyzing high-throughput sequence data for the DBA/2J strain. This study revealed that BXD progeny are segregating for considerable genetic variation including 4.46 million SNPs, 0.94 million indels, and 20,000 high confidence structural variants. The comprehensive annotation of these variants against mouse gene model (RefSeq) revealed that vast majority of them occur within intergenic (45%) and intronic (28%) regions. Around 40,000 variants occur within exons and a subset of which belongs to high-impact category including 61 nonsense, 196 splice-site, 99 frameshifts and 10,895 missense mutations. The high-impact coding variants that I uncovered provide an unprecedented resource with which to establish genome to phenome relations in a remarkably large and stable set of BXD lines.

In chapter 3, I explored the functional impact of genetic variation on gene expression by performing allele-specific gene expression (ASE) analysis using liver RNA-seq data from isogenic F1 hybrids. This study revealed that *cis* acting variation in expression is pervasive and is detected in roughly 50% of all assayable genes. Over one-third of them differ in expression greater than two-fold. This study also revealed that genes exhibiting strong ASE differences have a higher density of sequence variants, higher functional redundancy, and lower evolutionary conservation compared to genes with no ASE. Genes exhibiting high ASE differences in conjugation with high-impact coding variants (Chapter 2) should be key molecular resources for reverse genetics analysis.

In chapter 4, I exploited list of high-impact genetic variants (genes) that I generated in chapters two and three to perform a phenome-wide association study to investigate genome-to-phenome relations at multiple scales—from mRNA and protein levels to disease risk, behavior, and environmental interactions. We also exploited a large human clinical cohort—the Vanderbilt BioVU cohort— for validation and cross species translation of the novel associations. We successfully replicated almost all of the known genome-to-phenome associations in BXDs that were previously identified using forward genetics approach, and also identified a few novel associations. For example, we linked a missense mutation in fumarate hydratase that controls variation in the mitochondrial unfolded protein response in both mouse and *C. elegans*, and missense mutations in *Col6a5* that underlies variation in bone mineral density in both mouse and human. However, downstream effects of allelic variants with presumed deleterious effects on gene expression or protein structure are often small or undetectable. This may often be due to a lack of technical sensitivity and power or to phenotypic buffering. We demonstrate that phenome scans can be effective at linking sequence variants to a range of phenotypes and can be used to identify novel genome-to-phenome relations or validate hypothesized associations from independent studies. Coupling mouse and human PheWAS cohorts also shows great promise, and provides an efficient method to validate and translate key genome-to-phenome relations.

LIST OF REFERENCES

1. King RA, Rotter JI, Motulsky AG (1992) The Genetic basis of common diseases. New York: Oxford University Press. xiii, 978 p. p.
2. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
3. Falconer DS (1981) Introduction to quantitative genetics. London ; New York: Longman. viii, 340 p. p.
4. Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, Mass.: Sinauer. xvi, 980 p. p.
5. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature 526: 75-81.
6. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.
7. Kamide K, Kokubo Y, Yang J, Tanaka C, Hanada H, et al. (2005) Hypertension susceptibility genes on chromosome 2p24-p25 in a general Japanese population. J Hypertens 23: 955-960.
8. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, et al. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38: 617-619.
9. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
10. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43: D789-798.
11. Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. (2015) The UK10K project identifies rare variants in health and disease. Nature 526: 82-90.
12. Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, et al. (2014) Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet 10: e1004494.
13. Sulem P, Gudbjartsson DF, Walters GB, Helgadottir HT, Helgason A, et al. (2011) Identification of low-frequency variants associated with gout and serum uric acid levels. Nat Genet 43: 1127-1130.
14. Munker T (1999) [Phamarcogenomics: personalized drugs and personalized medicine]. J Pharm Belg 54: 125-129.
15. Ginsburg GS, McCarthy JJ (2001) Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol 19: 491-496.
16. Mancinelli L, Cronin M, Sadee W (2000) Pharmacogenomics: the promise of personalized medicine. AAPS PharmSci 2: E4.
17. Brock DC (2007) Understanding Moore's law: four decades of innovation. Choice: Current Reviews for Academic Libraries 44: 1944-1944.

18. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.
19. Flint J (2011) Mapping quantitative traits and strategies to find quantitative trait genes. Methods 53: 163-174.
20. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, et al. (1986) Cloning the gene for an inherited human disorder--chronic granulomatous disease--on the basis of its chromosomal location. Nature 322: 32-38.
21. Ray PN, Belfall B, Duff C, Logan C, Kean V, et al. (1985) Cloning of the breakpoint of an X;21 translocation associated with Duchenne muscular dystrophy. Nature 318: 672-675.
22. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245: 1066-1073.
23. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72: 971-983.
24. Goodfellow PN (1993) Planting alfalfa and cloning the Huntington's disease gene. Cell 72: 817-818.
25. Reeders ST, Breuning MH, Davies KE, Nicholls RD, Jarman AP, et al. (1985) A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. Nature 317: 542-544.
26. Reeders ST, Breuning MH, Corney G, Jeremiah SJ, Meera Khan P, et al. (1986) Two genetic markers closely linked to adult polycystic kidney disease on chromosome 16. Br Med J (Clin Res Ed) 292: 851-853.
27. Onuchic LF, Furu L, Nagasawa Y, Hou X, Eggermann T, et al. (2002) PKHD1, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple immunoglobulin-like plexin-transcription-factor domains and parallel beta-helix 1 repeats. Am J Hum Genet 70: 1305-1317.
28. Lidksy AS, Robson KJ, Thirumalachary C, Barker PE, Ruddle FH, et al. (1984) The PKU locus in man is on chromosome 12. Am J Hum Genet 36: 527-533.
29. Shibahara S, Tomita Y, Tagami H, Muller RM, Cohen T (1988) Molecular basis for the heterogeneity of human tyrosinase. Tohoku J Exp Med 156: 403-414.
30. Rubinsztein DC, Carmichael J (2003) Huntington's disease: molecular basis of neurodegeneration. Expert Rev Mol Med 5: 1-21.
31. Marian AJ (2012) Molecular genetic studies of complex phenotypes. Transl Res 159: 64-79.
32. (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature 511: 421-427.
33. Lander ES (1996) The new genomics: global views of biology. Science 274: 536-539.
34. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6: 109-118.
35. (2003) The International HapMap Project. Nature 426: 789-796.
36. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

37. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. Nature 526: 68-74.
38. Lee JY, Lee BS, Shin DJ, Woo Park K, Shin YA, et al. (2013) A genome-wide association study of a coronary artery disease risk variant. J Hum Genet 58: 120-126.
39. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39: 770-775.
40. Tamiya G, Shinya M, Imanishi T, Ikuta T, Makino S, et al. (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. Hum Mol Genet 14: 2305-2321.
41. Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, et al. (2008) Whole-genome association study of bipolar disorder. Mol Psychiatry 13: 558-569.
42. Von Wowern F, Bengtsson K, Lindblad U, Rastam L, Melander O (2004) Functional variant in the (alpha)2B adrenoceptor gene, a positional candidate on chromosome 2, associates with hypertension. Hypertension 43: 592-597.
43. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512-517.
44. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
45. Devlin B, Roeder K, Wasserman L (2000) Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. Biostatistics 1: 369-387.
46. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60: 155-166.
47. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004.
48. Hoffman GE (2013) Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One 8: e75707.
49. Steindel SJ (2010) International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. J Am Med Inform Assoc 17: 274-282.
50. Slee VN (1978) The International Classification of Diseases: ninth revision (ICD-9). Ann Intern Med 88: 424-426.
51. Bramer GR (1988) International statistical classification of diseases and related health problems. Tenth revision. World Health Stat Q 41: 32-36.
52. Denny JC, Miller RA, Waitman LR, Arrieta MA, Peterson JF (2009) Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. Int J Med Inform 78 Suppl 1: S34-42.
53. Pakhomov S, Bjornsen S, Hanson P, Smith S (2008) Quality performance measurement using the text of electronic medical records. Med Decis Making 28: 462-470.
54. Namjou B, Marsolo K, Caroll RJ, Denny JC, Ritchie MD, et al. (2014) Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically

links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. Front Genet 5: 401.

55. Ye Z, Mayer J, Ivacic L, Zhou Z, He M, et al. (2015) Phenome-wide association studies (PheWASs) for functional variants. Eur J Hum Genet 23: 523-529.
56. Shameer K, Denny JC, Ding K, Jouni H, Crosslin DR, et al. (2014) A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Hum Genet 133: 95-109.
57. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, et al. (2011) Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet 89: 529-542.
58. Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, et al. (2011) The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. Genet Epidemiol 35: 410-422.
59. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26: 1205-1210.
60. Clayton EW, Smith M, Fullerton SM, Burke W, McCarty CA, et al. (2010) Confronting real time ethical, legal, and social issues in the Electronic Medical Records and Genomics (eMERGE) Consortium. Genet Med 12: 616-620.
61. Fullerton SM, Wolf WA, Brothers KB, Clayton EW, Crawford DC, et al. (2012) Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network. Genet Med 14: 424-431.
62. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR (2010) Principles of human subjects protections applied in an opt-out, de-identified biobank. Clin Transl Sci 3: 42-48.
63. McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, et al. (2013) Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: pediatric BioVU. Clin Pharmacol Ther 93: 204-211.
64. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, et al. (2013) Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet 9: e1003087.
65. Pathak J, Kiefer RC, Bielinski SJ, Chute CG (2012) Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank. J Biomed Semantics 3: 10.
66. Hebbring SJ, Schrodi SJ, Ye Z, Zhou Z, Page D, et al. (2013) A PheWAS approach in studying HLA-DRB1*1501. Genes Immun 14: 187-191.
67. Hall MA, Verma A, Brown-Gentry KD, Goodloe R, Boston J, et al. (2014) Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. PLoS Genet 10: e1004678.

68. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, et al. (2012) Use of genome-wide association studies for drug repositioning. Nat Biotechnol 30: 317-320.
69. Schnabel RB, Yin X, Gona P, Larson MG, Beiser AS, et al. (2015) 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. Lancet.
70. Wilcox M, Li Q, Sun Y, Stang P, Berlin J, et al. (2009) Genome-wide association study for empirically derived metabolic phenotypes in the Framingham Heart Study offspring cohort. BMC Proc 3 Suppl 7: S53.
71. Kreger BE, Splansky GL, Schatzkin A (1991) The cancer experience in the Framingham Heart Study cohort. Cancer 67: 1-6.
72. Moscicki EK, Elkins EF, Baum HM, McNamara PM (1985) Hearing loss in the elderly: an epidemiologic study of the Framingham Heart Study Cohort. Ear Hear 6: 184-190.
73. Phillips GB, Castelli WP, Abbott RD, McNamara PM (1983) Association of hyperestrogenemia and coronary heart disease in men in the Framingham cohort. Am J Med 74: 863-869.
74. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, et al. (2013) Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 42: 111-127.
75. Ziebarth JD, Cook MN, Wang X, Williams RW, Lu L, et al. (2012) Treatment- and population-dependent activity patterns of behavioral and expression QTLs. PLoS One 7: e31805.
76. Austin MA, Hair MS, Fullerton SM (2012) Research guidelines in the era of large-scale collaborations: an analysis of Genome-wide Association Study Consortia. Am J Epidemiol 175: 962-969.
77. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, et al. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38: 879-887.
78. Taylor BA, Wnek C, Kotlus BS, Roemer N, MacTaggart T, et al. (1999) Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. Mamm Genome 10: 335-348.
79. Williams RW, Bennett B, Lu L, Gu J, DeFries JC, et al. (2004) Genetic structure of the LXS panel of recombinant inbred mouse strains: a powerful resource for complex trait analysis. Mamm Genome 15: 637-647.
80. Threadgill DW, Miller DR, Churchill GA, de Villena FP (2011) The collaborative cross: a recombinant inbred mouse population for the systems genetic era. ILAR J 52: 24-31.
81. Flint J, Eskin E (2012) Genome-wide association studies in mice. Nat Rev Genet 13: 807-817.
82. Pandey AK, Williams RW (2014) Genetics of gene expression in CNS. Int Rev Neurobiol 116: 195-231.
83. Williams RW, Gu J, Qi S, Lu L (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. Genome Biol 2: RESEARCH0046.

84. Taylor BA, Heiniger HJ, Meier H (1973) Genetic analysis of resistance to cadmium-induced testicular damage in mice. Proc Soc Exp Biol Med 143: 629-633.

85. Koutnikova H, Laakso M, Lu L, Combe R, Paananen J, et al. (2009) Identification of the UBP1 locus as a critical blood pressure determinant using a combination of mouse and human genetics. PLoS Genet 5: e1000591.

86. Andreux PA, Williams EG, Koutnikova H, Houtkooper RH, Champy MF, et al. (2012) Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. Cell 150: 1287-1299.

87. Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, et al. (2013) Mitonuclear protein imbalance as a conserved longevity mechanism. Nature 497: 451-457.

88. Li Z, Mulligan MK, Wang X, Miles MF, Lu L, et al. (2010) A transposon in Comt generates mRNA variants and causes widespread expression and behavioral differences among mice. PLoS One 5: e12181.

89. Wang X, Mozhui K, Li Z, Mulligan MK, Ingels FJ, et al. (2012) A promoter polymorphism in the Per3 gene is associated with alcohol and stress response. Transl Psychiatry 2: 81-85.

90. Lee JC, Lyons PA, McKinney EF, Sowerby JM, Carr EJ, et al. (2011) Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. J Clin Invest 121: 4170-4179.

91. Kim WJ, Kim SK, Jeong P, Yun SJ, Cho IC, et al. (2011) A four-gene signature predicts disease progression in muscle invasive bladder cancer. Mol Med 17: 478-485.

92. Sotiriou C, Piccart MJ (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7: 545-553.

93. Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, et al. (2006) How replicable are mRNA expression QTL? Mamm Genome 17: 643-656.

94. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

95. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190-1195.

96. Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 30: 1095-1106.

97. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302.

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

99. Cohen BB, Porter DE, Wallace MR, Carothers A, Steel CM (1993) Linkage of a major breast cancer gene to chromosome 17q12-21: results from 15 Edinburgh families. Am J Hum Genet 52: 723-729.

100. Albertsen HM, Smith SA, Mazoyer S, Fujimoto E, Stevens J, et al. (1994) A physical map and candidate genes in the BRCA1 region on chromosome 17q12-21. Nat Genet 7: 472-479.

101. Porter DE, Cohen BB, Wallace MR, Smyth E, Chetty U, et al. (1994) Breast cancer incidence, penetrance and survival in probable carriers of BRCA1 gene mutation in families linked to BRCA1 on chromosome 17q12-21. Br J Surg 81: 1512-1515.

102. Oddoux C, Struewing JP, Clayton CM, Neuhausen S, Brody LC, et al. (1996) The carrier frequency of the BRCA2 6174delT mutation among Ashkenazi Jewish individuals is approximately 1%. Nat Genet 14: 188-190.

103. Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, et al. (1994) Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature 368: 258-261.

104. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, et al. (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 75: 1027-1038.

105. Verma L, Kane MF, Brassett C, Schmeits J, Evans DG, et al. (1999) Mononucleotide microsatellite instability and germline MSH6 mutation analysis in early onset colorectal cancer. J Med Genet 36: 678-682.

106. Nicolaides NC, Papadopoulos N, Liu B, Wei YF, Carter KC, et al. (1994) Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. Nature 371: 75-80.

107. Kamb A, Gruis NA, Weaver-Feldhaus J, Liu Q, Harshman K, et al. (1994) A cell cycle regulator potentially involved in genesis of many tumor types. Science 264: 436-440.

108. Winchester RJ (1981) Genetic aspects of rheumatoid arthritis. Springer Semin Immunopathol 4: 89-102.

109. Dahlback B (1994) Inherited resistance to activated protein C, a major cause of venous thrombosis, is due to a mutation in the factor V gene. Haemostasis 24: 139-151.

110. Koster T, Rosendaal FR, Reitsma PH, van der Velden PA, Briet E, et al. (1994) Factor VII and fibrinogen levels as risk factors for venous thrombosis. A case-control study of plasma levels and DNA polymorphisms--the Leiden Thrombophilia Study (LETS). Thromb Haemost 71: 719-722.

111. Arruda VR, von Zuben PM, Chiaparini LC, Annichino-Bizzacchi JM, Costa FF (1997) The mutation Ala677-->Val in the methylene tetrahydrofolate reductase gene: a risk factor for arterial disease and venous thrombosis. Thromb Haemost 77: 818-821.

112. Pruthi S, Gostout BS, Lindor NM (2010) Identification and Management of Women With BRCA Mutations or Hereditary Predisposition for Breast and Ovarian Cancer. Mayo Clin Proc 85: 1111-1120.

113. Trotter TL, Fleischman AR, Howell RR, Lloyd-Puryear M (2011) Secretary's Advisory Committee on Heritable Disorders in Newborns and Children response to the President's Council on Bioethics report: the changing moral focus of newborn screening. Genet Med 13: 301-304.

114. Ingelman-Sundberg M (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J 5: 6-13.

115. Kirchheiner J, Schmidt H, Tzvetkov M, Keulen JT, Lotsch J, et al. (2007) Pharmacokinetics of codeine and its metabolite morphine in ultra-rapid metabolizers due to CYP2D6 duplication. Pharmacogenomics J 7: 257-265.
116. Minuti G, D'Incecco A, Cappuzzo F (2013) Targeted therapy for NSCLC with driver mutations. Expert Opin Biol Ther 13: 1401-1412.
117. Duffy MJ, Crown J (2013) Companion biomarkers: paving the pathway to personalized treatment for cancer. Clin Chem 59: 1447-1456.
118. Flaherty KT, Robert C, Hersey P, Nathan P, Garbe C, et al. (2012) Improved survival with MEK inhibition in BRAF-mutated melanoma. N Engl J Med 367: 107-114.
119. Janne PA, Shaw AT, Pereira JR, Jeannin G, Vansteenkiste J, et al. (2013) Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. Lancet Oncol 14: 38-47.
120. Bonfield JK, Mahoney MV (2013) Compression of FASTQ and SAM format sequencing data. PLoS One 8: e59190.
121. Campagne F, Dorff KC, Chambwe N, Robinson JT, Mesirov JP (2013) Compression of structured high-throughput sequencing data. PLoS One 8: e79871.
122. Pyl PT, Gehring J, Fischer B, Huber W (2014) h5vc: scalable nucleotide tallies with HDF5. Bioinformatics 30: 1464-1466.
123. Mason CE, Zumbo P, Sanders S, Folk M, Robinson D, et al. (2010) Standardizing the next generation of bioinformatics software development with BioHDF (HDF5). Adv Exp Med Biol 680: 693-700.
124. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 21: 734-740.
125. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6: e1001025.
126. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-913.
127. Chiang CW, Liu CT, Lettre G, Lange LA, Jorgensen NW, et al. (2012) Ultraconserved elements in the human genome: association and transmission analyses of highly constrained single-nucleotide polymorphisms. Genetics 192: 253-266.
128. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. (2007) Human genome ultraconserved elements are ultraselected. Science 317: 915.
129. Chen CT, Wang JC, Cohen BA (2007) The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet 80: 692-704.
130. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321-1325.
131. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, et al. (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 42: D142-147.

132. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036-1040.
133. Gershenzon NI, Stormo GD, Ioshikhes IP (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res 33: 2290-2301.
134. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.
135. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46: 310-315.
136. Ritchie GR, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nat Methods 11: 294-296.
137. Hsu YH, Zillikens MC, Wilson SG, Farber CR, Demissie S, et al. (2010) An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. PLoS Genet 6: e1000977.
138. Zhang M, Liang L, Morar N, Dixon AL, Lathrop GM, et al. (2012) Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. Hum Genet 131: 615-623.
139. Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, et al. (2010) Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet 6: e1000932.
140. Huang YT, Vanderweele TJ, Lin X (2014) Joint Analysis of Snp and Gene Expression Data in Genetic Association Studies of Complex Diseases. Ann Appl Stat 8: 352-376.
141. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, et al. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 37: 225-232.
142. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, et al. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature Genet 37: 233-242.
143. Wang J, Kong L, Gao G, Luo J (2013) A brief introduction to web-based genome browsers. Brief Bioinform 14: 131-143.
144. Anderson MG, Smith RS, Hawes NL, Zabaleta A, Chang B, et al. (2002) Mutations in genes encoding melanosomal proteins cause pigmentary glaucoma in DBA/2J mice. Nat Genet 30: 81-85.
145. Chang B, Smith RS, Hawes NL, Anderson MG, Zabaleta A, et al. (1999) Interacting loci cause severe iris atrophy and glaucoma in DBA/2J mice. Nat Genet 21: 405-409.
146. Howell GR, Libby RT, Marchant JK, Wilson LA, Cosma IM, et al. (2007) Absence of glaucoma in DBA/2J mice homozygous for wild-type versions of Gpnmb and Tyrp1. BMC Genet 8: 45.

147. Someya S, Yamasoba T, Prolla TA, Tanokura M (2007) Genes encoding mitochondrial respiratory chain components are profoundly down-regulated with aging in the cochlea of DBA/2J mice. Brain Res 1182: 26-33.

148. Johnson KR, Longo-Guess C, Gagnon LH, Yu H, Zheng QY (2008) A locus on distal chromosome 11 (ahl8) and its interaction with Cdh23 ahl underlie the early onset, age-related hearing loss of DBA/2J mice. Genomics 92: 219-225.

149. Paigen B, Morrow A, Brandon C, Mitchell D, Holmes P (1985) Variation in susceptibility to atherosclerosis among inbred strains of mice. Atherosclerosis 57: 65-73.

150. Nishina PM, Wang J, Toyofuku W, Kuypers FA, Ishida BY, et al. (1993) Atherosclerosis and plasma and liver lipids in nine inbred strains of mice. Lipids 28: 599-605.

151. Davis RC, Schadt EE, Cervino AC, Peterfy M, Lusis AJ (2005) Ultrafine mapping of SNPs from mouse strains C57BL/6J, DBA/2J, and C57BLKS/J for loci contributing to diabetes and atherosclerosis susceptibility. Diabetes 54: 1191-1199.

152. Broadbent J, Kampmueller KM, Koonse SA (2005) Role of dopamine in behavioral sensitization to ethanol in DBA/2J mice. Alcohol 35: 137-148.

153. Berrettini WH, Alexander R, Ferraro TN, Vogel WH (1994) A study of oral morphine preference in inbred mouse strains. Psychiatr Genet 4: 81-86.

154. Belknap JK, Crabbe JC, Riggan J, O'Toole LA (1993) Voluntary consumption of morphine in 15 inbred mouse strains. Psychopharmacology (Berl) 112: 352-358.

155. Belknap JK, Noordewier B, Lame M (1989) Genetic dissociation of multiple morphine effects among C57BL/6J, DBA/2J and C3H/HeJ inbred mouse strains. Physiol Behav 46: 69-74.

156. Frigeni V, Bruno F, Carenzi A, Racagni G (1981) Difference in the development of tolerance to morphine and D-ALA2-methionine-enkephalin in C57 BL/6J and DBA/2J mice. Life Sci 28: 729-736.

157. Belknap JK, Crabbe JC, Young ER (1993) Voluntary consumption of ethanol in 15 inbred mouse strains. Psychopharmacology (Berl) 112: 503-510.

158. Risinger FO, Cunningham CL (1995) Genetic differences in ethanol-induced conditioned taste aversion after ethanol preexposure. Alcohol 12: 535-539.

159. Trammell RA, Liberati TA, Toth LA (2012) Host genetic background and the innate inflammatory response of lung to influenza virus. Microbes Infect 14: 50-58.

160. Boon AC, deBeauchamp J, Hollmann A, Luke J, Kotb M, et al. (2009) Host genetic variation affects resistance to infection with a highly pathogenic H5N1 influenza A virus in mice. J Virol 83: 10417-10426.

161. Srivastava B, Blazejewska P, Hessmann M, Bruder D, Geffers R, et al. (2009) Host genetic background strongly influences the response to influenza a virus infections. PLoS One 4: e4857.

162. Zhang Y, Huang J, Jiao Y, David V, Kocak M, et al. (2015) Bone morphology in 46 BXD recombinant inbred strains and femur-tibia correlation. ScientificWorldJournal 2015: 728278.

163. Bower AL, Lang DH, Vogler GP, Vandenbergh DJ, Blizard DA, et al. (2006) QTL analysis of trabecular bone in BXD F2 and RI mice. J Bone Miner Res 21: 1267-1275.
164. Akhter MP, Fan Z, Rho JY (2004) Bone intrinsic material properties in three inbred mouse strains. Calcif Tissue Int 75: 416-420.
165. Beamer WG, Donahue LR, Rosen CJ, Baylink DJ (1996) Genetic variability in adult bone density among inbred strains of mice. Bone 18: 397-403.
166. Bourson A, Kapps V, Zwingelstein C, Rudler A, Boess FG, et al. (1997) Correlation between 5-HT7 receptor affinity and protection against sound-induced seizures in DBA/2J mice. Naunyn Schmiedebergs Arch Pharmacol 356: 820-826.
167. Banko ML, Allen KM, Dolina S, Neumann PE, Seyfried TN (1997) Genomic imprinting and audiogenic seizures in mice. Behav Genet 27: 465-475.
168. Marchand MJ, Ward R, Moreau B (1996) Different patterns of dendritic branching of posterior collicular neurons in two mouse strains are associated with a difference in audiogenic seizure susceptibility. J Hirnforsch 37: 135-143.
169. Schreiber RA (1982) Effect of ethanol on audiogenic seizures in C57BL/6J and DBA/2J mice. Subst Alcohol Actions Misuse 3: 325-330.
170. Taylor BA (1972) Genetic relationships between inbred strains of mice. J Hered 63: 83-86.
171. Carneiro AM, Airey DC, Thompson B, Zhu CB, Lu L, et al. (2009) Functional coding variation in recombinant inbred mouse lines reveals multiple serotonin transporter-associated phenotypes. Proc Natl Acad Sci U S A 106: 2047-2052.
172. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, et al. (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477: 289-294.
173. Wang X, Agarwala R, Capra J, Chen Z, Church D, et al. (2010) High-throughput sequencing of the DBA/2J mouse genome. BMC Bioinformatics 11: O7.
174. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.
175. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
176. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
177. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
178. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871.
179. Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinformatics 2014.
180. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6: 677-681.

181. Simpson JT, McIntyre RE, Adams DJ, Durbin R (2010) Copy number variant detection in inbred strains from short read sequence data. Bioinformatics 26: 565-567.

182. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31: 51-54.

183. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.

184. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7: Unit7 20.

185. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115-119.

186. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, et al. (2002) The Pfam protein families database. Nucleic Acids Res 30: 276-280.

187. Mozhui K, Ciobanu DC, Schikorski T, Wang X, Lu L, et al. (2008) Dissection of a QTL hotspot on mouse distal chromosome 1 that modulates neurobehavioral phenotypes and gene expression. PLoS Genet 4: e1000260.

188. Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, et al. (2011) Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43: 648-655.

189. Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F (2007) On the subspecific origin of the laboratory mouse. Nat Genet 39: 1100-1107.

190. Chang C, Smith DR, Prasad VS, Sidman CL, Nebert DW, et al. (1993) Ten nucleotide differences, five of which cause amino acid changes, are associated with the Ah receptor locus polymorphism of C57BL/6 and DBA/2 mice. Pharmacogenetics 3: 312-321.

191. del Pilar Jimenez AM, Viriyakosol S, Walls L, Datta SK, Kirkland T, et al. (2008) Susceptibility to Coccidioides species in C57BL/6 mice is associated with expression of a truncated splice variant of Dectin-1 (Clec7a). Genes Immun 9: 338-348.

192. Wetsel RA, Fleischer DT, Haviland DL (1990) Deficiency of the murine fifth complement component (C5). A 2-base pair gene deletion in a 5'-exon. J Biol Chem 265: 2435-2440.

193. Pasinetti GM, Tocco G, Sakhi S, Musleh WD, DeSimoni MG, et al. (1996) Hereditary deficiencies in complement C5 are associated with intensified neurodegenerative responses that implicate new roles for the C-system in neuronal and astrocytic functions. Neurobiol Dis 3: 197-204.

194. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217-1224.

195. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature 488: 116-120.

196. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281-297.

197. Del Punta K, Leinders-Zufall T, Rodriguez I, Jukam D, Wysocki CJ, et al. (2002) Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. Nature 419: 70-74.
198. Williams Rt, Lim JE, Harr B, Wing C, Walters R, et al. (2009) A common and unstable copy number variant is associated with differences in Glo1 expression and anxiety-like behavior. PLoS One 4: e4649.
199. Distler MG, Palmer AA (2012) Role of Glyoxalase 1 (Glo1) and methylglyoxal (MG) in behavior: recent advances and mechanistic insights. Front Genet 3: 250.
200. Marshall E (2001) Genome sequencing. Celera assembles mouse genome; public labs plan new strategy. Science 292: 822.
201. Jiang Y, Turinsky AL, Brudno M (2015) The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Res.
202. Damerval C, Maurice A, Josse JM, de Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 137: 289-301.
203. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752-755.
204. Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, et al. (2015) Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nat Genet 47: 353-360.
205. Lonsdale J, Thomas J, Salvatore M (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580-585.
206. Keen JC, Moore HM (2015) The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. J Pers Med 5: 22-29.
207. Ciobanu DC, Lu L, Mozhui K, Wang X, Jagalur M, et al. (2010) Detection, validation, and downstream analysis of allelic variation in gene expression. Genetics 184: 119-128.
208. Bell GD, Kane NC, Rieseberg LH, Adams KL (2013) RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. Genome Biol Evol 5: 1309-1323.
209. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, et al. (2010) Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res 20: 816-825.
210. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, et al. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol 7: 522.
211. Szabo PE, Mann JR (1995) Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanisms. Genes Dev 9: 3097-3108.
212. DeVeale B, van der Kooy D, Babak T (2012) Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS Genet 8: e1002600.

213. Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, et al. (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One 3: e3839.

214. Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, et al. (2013) Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage. Genetics 195: 1157-1166.

215. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

216. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57: 289-300.

217. Geisert EE, Lu L, Freeman-Anderson NE, Templeton JP, Nassr M, et al. (2009) Gene expression in the mouse eye: an online resource for genetics using 103 strains of mice. Mol Vis 15: 1730-1763.

218. Haley CS, Knott SA, Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136: 1195-1207.

219. Wu Y, Williams EG, Dubuis S, Mottis A, Jovaisaite V, et al. (2014) Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell 158: 1415-1430.

220. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

221. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal--unified access to biological data. Nucleic Acids Res 37: W23-27.

222. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327-335.

223. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207-3212.

224. Satya RV, Zavaljevski N, Reifman J (2012) A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Res 40: e127.

225. Hilgers V, Perry MW, Hendrix D, Stark A, Levine M, et al. (2011) Neural-specific elongation of 3' UTRs during Drosophila development. Proc Natl Acad Sci U S A 108: 15864-15869.

226. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC (2013) Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res 23: 812-825.

227. Lin S, Yang Z, Liu H, Cai Z (2011) Metabolomic analysis of liver and skeletal muscle tissues in C57BL/6J and DBA/2J mice exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin. Mol Biosyst 7: 1956-1965.

228. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13: 418.

229. Daily K, Patel VR, Rigor P, Xie X, Baldi P (2011) MotifMap: integrative genome-wide maps of regulatory motif sites for model species. BMC Bioinformatics 12: 495.

230. Eisenberg E, Levanon EY (2013) Human housekeeping genes, revisited. Trends Genet 29: 569-574.

231. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

232. Weng MP, Liao BY (2010) MamPhEA: a web tool for mammalian phenotype enrichment analysis. Bioinformatics 26: 2212-2213.

233. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE (2009) The Mouse Genome Database genotypes::phenotypes. Nucleic Acids Res 37: D712-719.

234. Baker KE, Parker R (2004) Nonsense-mediated mRNA decay: terminating erroneous gene expression. Curr Opin Cell Biol 16: 293-299.

235. Zhang Z, Xin D, Wang P, Zhou L, Hu L, et al. (2009) Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC Biol 7: 23.

236. Lareau LF, Brooks AN, Soergel DA, Meng Q, Brenner SE (2007) The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv Exp Med Biol 623: 190-211.

237. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.

238. Liang H, Cavalcanti AR, Landweber LF (2005) Conservation of tandem stop codons in yeasts. Genome Biol 6: R31.

239. An JJ, Gharami K, Liao GY, Woo NH, Lau AG, et al. (2008) Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. Cell 134: 175-187.

240. Kendig EL, Chen Y, Krishan M, Johansson E, Schneider SN, et al. (2011) Lipid metabolism and body composition in Gclm(-/-) mice. Toxicol Appl Pharmacol 257: 338-348.

241. Aston-Mourney K, Wong N, Kebede M, Zraika S, Balmer L, et al. (2007) Increased nicotinamide nucleotide transhydrogenase levels predispose to insulin hypersecretion in a mouse strain susceptible to diabetes. Diabetologia 50: 2476-2485.

242. Lu A, Wangpu X, Han D, Feng H, Zhao J, et al. (2012) TXNDC9 expression in colorectal cancer cells and its influence on colorectal cancer prognosis. Cancer Invest 30: 721-726.

243. Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11: R90.

244. Busque SM, Wagner CA (2009) Potassium restriction, high protein intake, and metabolic acidosis increase expression of the glutamine transporter SNAT3 (Slc38a3) in mouse kidney. Am J Physiol Renal Physiol 297: F440-450.

245. Busque SM, Stange G, Wagner CA (2014) Dysregulation of the glutamine transporter Slc38a3 (SNAT3) and ammoniagenic enzymes in obese, glucose-intolerant mice. Cell Physiol Biochem 34: 575-589.

246. Mitterberger MC, Kim G, Rostek U, Levine RL, Zwerschke W (2012) Carbonic anhydrase III regulates peroxisome proliferator-activated receptor-gamma2. Exp Cell Res 318: 877-886.

247. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21: 236-239.

248. She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, et al. (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics 10: 269.

249. Masel J, Siegal ML (2009) Robustness: mechanisms and consequences. Trends Genet 25: 395-403.

250. Waddington CH (1942) Canalization of development and the inheritance of acquired characters. Nature 150: 563.

251. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, et al. (2011) Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet 7: e1001393.

252. Albert FW, Treusch S, Shockley AH, Bloom JS, Kruglyak L (2014) Genetics of single-cell protein abundance variation in large yeast populations. Nature 506: 494-497.

253. Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, et al. (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res 23: 1496-1504.

254. Maier T, Guell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett 583: 3966-3973.

255. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, et al. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genet 41: 166-167.

256. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global quantification of mammalian gene expression control. Nature 473: 337-342.

257. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2: 65-73.

258. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21: 2213-2223.

259. Babak T, Garrett-Engele P, Armour CD, Raymond CK, Keller MP, et al. (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. BMC Genomics 11: 473.

260. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, et al. (2014) Allele-specific expression and eQTL analysis in mouse adipose tissue. BMC Genomics 15: 471.

261. Mozhui K, Lu L, Armstrong WE, Williams RW (2012) Sex-specific modulation of gene expression networks in murine hypothalamus. Front Neurosci 6: 63.

262. Williams EG, Mouchiroud L, Frochaux M, Pandey A, Andreux PA, et al. (2014) An evolutionarily conserved role for the aryl hydrocarbon receptor in the regulation of movement. PLoS Genet 10: e1004673.

263. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228-237.

264. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628-640.

265. Jelier R, Semple JI, Garcia-Verdugo R, Lehner B (2011) Predicting phenotypic variation in yeast from individual genome sequences. Nat Genet 43: 1270-1274.

266. Bogani D, Warr N, Elms P, Davies J, Tymowska-Lalanne Z, et al. (2004) New semidominant mutations that affect mouse development. Genesis 40: 109-117.

267. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, et al. (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 31: 1102-1110.

268. Chesler EJ, Wang J, Lu L, Qu Y, Manly KF, et al. (2003) Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. Neuroinformatics 1: 343-357.

269. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. Nat Neurosci 7: 485-486.

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

271. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

272. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

273. Guidry PA, Stroynowski I (2005) The murine family of gut-restricted class Ib MHC includes alternatively spliced isoforms of the proposed HLA-G homolog, "blastocyst MHC". J Immunol 175: 5248-5259.

274. Tomlinson IP, Alam NA, Rowan AJ, Barclay E, Jaeger EE, et al. (2002) Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. Nat Genet 30: 406-410.

275. Sena LA, Li S, Jairaman A, Prakriya M, Ezponda T, et al. (2013) Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. Immunity 38: 225-236.

276. Gara SK, Grumati P, Urciuolo A, Bonaldo P, Kobbe B, et al. (2008) Three novel collagen VI chains with high homology to the alpha3 chain. J Biol Chem 283: 10658-10670.

277. Ema M, Ohe N, Suzuki M, Mimura J, Sogawa K, et al. (1994) Dioxin binding activities of polymorphic forms of mouse and human arylhydrocarbon receptors. J Biol Chem 269: 27337-27343.

278. Runkel ED, Liu S, Baumeister R, Schulze E (2013) Surveillance-activated defenses block the ROS-induced mitochondrial unfolded protein response. PLoS Genet 9: e1003346.

279. Zhao Q, Wang J, Levichkin IV, Stasinopoulos S, Ryan MT, et al. (2002) A mitochondrial specific stress response in mammalian cells. EMBO J 21: 4411-4419.

280. Durieux J, Wolff S, Dillin A (2011) The cell-non-autonomous nature of electron transport chain-mediated longevity. Cell 144: 79-91.

281. McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, et al. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proceedings of the National Academy of Sciences of the United States of America 107: 6544-6549.

282. Yamamoto S, Jaiswal M, Charng WL, Gambin T, Karaca E, et al. (2014) A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. Cell 159: 200-214.

283. Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, et al. (2014) Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Research 24: 1193-1208.

284. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, et al. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics 41: 166-167.

285. Klassen T, Davis C, Goldman A, Burgess D, Chen T, et al. (2011) Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell 145: 1036-1048.

286. Manu, Surkova S, Spirov AV, Gursky VV, Janssens H, et al. (2009) Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. PLoS Biology 7: e1000049.

287. Barbaric I, Miller G, Dear TN (2007) Appearances can be deceiving: phenotypes of knockout mice. Briefings in Functional Genomics and Proteomics 6: 91-103.

288. Hartman JLt, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. Science 291: 1001-1004.

VITA

Ashutosh K. Pandey was born in Damoh, India in the year 1985. He received Bachelor of Engineering in Biotechnology from Madhav Institute of Technology and Science, India, in July 2007. He then studied Bioinformatics at the University of Memphis, Memphis and received Master of Science in December, 2009. He was matriculated in Integrated Program in Biomedical Sciences for doctoral studies at the University of Tennessee Health Science Center, Memphis in August 2010. He received his doctorate degree in December 2015.