UNIVERSITY OF CALIFORNIA SAN DIEGO

Interrogating the effect of variation in *trans*-acting factors

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

in

Bioinformatics and Systems Biology

by

Cynthia Wu

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Alon Goren, Co-Chair
Professor Hannah Carter
Professor Bruce Hamilton
Professor Sven Heinz
Professor Pejman Mohammadi

2023

The Dissertation of Cynthia Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents who has supported me unconditionally since the beginning

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Table of Contents v

List of Figures vii

Acknowledgements ix

Vita xi

Abstract of the Dissertation xii

Introduction 1

Chapter 1 A novel quantitative trait locus implicates Msh3 in the propensity for genome-wide short tandem repeat expansions in mice 7

1.1 Abstract 7

1.2 Introduction 8

1.3 Results 10

1.3.1 Identifying new mutations in the BXD family 10

1.3.2 Mapping quantitative trait loci for STR mutation phenotypes 12

1.3.3 Analysis of candidate variants disrupting protein-coding genes 20

1.3.4 Expansion propensity QTL colococalizes with multiple cis-eQTLs 22

1.4 Discussion 27

1.5 Methods 34

1.5.1 WGS and variant calling in the BXD cohort 34

1.5.2 SNP marker maps for founder inference and interval mapping 36

1.5.3 Identifying and phasing new STR mutations 36

1.5.4 Inferring missing founder STR genotypes 37

1.5.5 Characterization of new STR mutations 37

1.5.6 Computing STR mutator phenotypes 38

1.5.7 QTL mapping for STR mutator phenotypes 38

1.5.8 Variant annotation 38

1.5.9 Pangenome analysis of SVs 39

1.5.10 eQTL analysis 40

1.5.11 Genomic data for classic mouse strains 43

1.5.12 Tissue-specific expression of DNA repair genes 43

1.6 Data access 43

1.7 Supplementary Figures 45

1.8 Acknowledgments 63

Chapter 2 xQTL: A mixture-model for identification and characterization of trans-eQTLs 65
2.1 Abstract 65
2.2 Introduction 66
2.3 Results 68
2.3.1 xQTL identifies trans-eQTLs with global effects 68
2.3.2 PEER correction removes true trans-eQTL effects 72
2.3.3 Validation of xQTL performance using yeast dataset 74
2.3.4 Genome-wide detection of trans-eQTLs in outbred rats 77
2.4 Discussion 82
2.5 Methods 83
2.5.1 Overview 83
2.5.2 Linear model for a single gene-by-variant pair 83
2.5.3 xQTL model 85
2.5.4 Fitting the mixture model 86
2.5.5 Simulation framework 87
2.5.6 Simulated datasets 87
2.5.7 Power analysis 89
2.5.8 Yeast dataset 89
2.5.9 Rat dataset 89
2.5.10 Trans-eQTL identification 90
2.6 Supplementary Figures 92
2.7 Acknowledgments 94

Chapter 3 scBE-seq: A pooled, high-precision genome editing strategy with single-cell sequencing to interrogate effects of hundreds of variants 95
3.1 Introduction 95
3.2 Results 97
3.2.1 Efficient base-editing of target mutations in HEK293T cells 97
3.2.2 Types of mutations scBE-seq can interrogate 100
3.2.3 Script to design gRNAs 101
3.3 Discussion 101
3.4 Acknowledgments 102

Chapter 4 Conclusions 103

References 105

LIST OF FIGURES

Figure 1.1. Characterizing new mutations in the BXD family 13

Figure 1.2. Discovery of QTLs for STR mutation phenotypes 15

Figure 1.3. Variants predicted to impact Msh3 23

Figure 1.4. The Chr 13 expansion propensity QTL colocalizes with eQTLs for multiple DNA repair genes 26

Figure 1.5. Schematic overview of proposed mechanisms for the expansion propensity QTL 31

Figure 1.6. Localization of new autosomal mutations at STRs 45

Figure 1.7. Summary of new STR mutations in BXD 46

Figure 1.8. Distribution of the number of strains carrying the new allele at each of the STRs for which at least one new mutation was identified 47

Figure 1.9. Expansion propensity phenotype at the Chr 13 QTL for mutations in genic vs. intergenic regions. 47

Figure 1.10. Genes located in or near the modest QTL peaks for expansion propensity . 48

Figure 1.11. Evaluating robustness of the Chr 13 association signal for expansion propensity 49

Figure 1.12. Distribution of expansion propensity for each strain for different repeat classes 50

Figure 1.13. Expansion propensity QTL mapping in each BXD epoch 50

Figure 1.14. Overview of STR mutation patterns 51

Figure 1.15. Mutation patterns stratified by repeat unit and haplotype at the Chr 13 QTL peak 52

Figure 1.16. Mutation patterns by haplotype at the Chr 13 QTL peak for autosomes, Chr X, and Chr Y 53

Figure 1.17. Features of the SVs discovered from pangenome analysis of Chr 13 53

Figure 1.18. Annotation and selection of impactful variants within genes in the Chr 13 QTL for expansion propensity 54

Figure 1.19. Detailed view of annotated variants within Msh3 55

Figure 1.20. Overall gene expression levels for genes within the QTL region 56

Figure 1.21. Summary of expansion propensity QTL signal detection using strains available in each gene expression dataset 57

Figure 1.22. Availability of gene expression data for genes within the expansion propensity QTL 57

Figure 1.23. Probe-level analysis of eQTL signals at Msh3 58

Figure 1.24. Summary of gene eQTL signals for genes contained within the QTL peak 1.5-LOD support interval for the expansion propensity phenotype 59

Figure 1.25. eQTL signals for DNA repair genes within the expansion propensity QTL 60

Figure 1.26. Co-localization of eQTL and expansion propensity signals 61

Figure 1.27. Detailed analysis of eQTL signals at Msh3. 62

Figure 1.28. Visualization of next-generation sequencing data for classic mouse strains at the 5’ end of Msh3 62

Figure 2.1. Simulation heatmap to benchmark xQTL 70

Figure 2.2. Tradeoff of trans-eQTL analyses with or without PC adjustment. 73

Figure 2.3. Application of xQTL, CPMA, and Matrix eQTL on yeast data 75

Figure 2.4. Cisregress removes the effects of cis-eQTLs 78

Figure 2.5. xQTL reveals trans-eQTL candidates 80

Figure 2.6. Overview of top-scoring trans-eQTL candidates 81

Figure 2.7. Simulation heatmap of xQTL vs CPMA 92

Figure 2.8. Empirical null accounts for gene correlation 93

Figure 3.1. Efficiencies of successful edits using pooled base editing in HEK293 98

Figure 3.2. Off-target editing of RNA 99

Figure 3.3. Heatmap of clusters from simulated scRNA-seq data 100

ACKNOWLEDGEMENTS

I am indebted to my advisors Melissa Gymrek and Alon Goren for their invaluable guidance and support. It has been an honor to learn from their extensive knowledge and vast research experience. Their advice and supervision were immensely helpful in developing and maturing my scientific identity.

I would like to thank my dissertation committee for their help and encouragement. Hannah Carter, Bruce Hamilton, Sven Heinz, and Pejman Mohammadi have all given me crucial feedback on my project in their respective fields. I would also like to extend my sincere thanks to Abraham Palmer whose important collaboration has helped shape my PhD projects and Rahel Wachs for help with figure illustrations.

I am grateful for my colleagues in the Gymrek lab and Goren lab for their insightful discussion and critiques. I very much appreciate all my collaborators who has offered incredible ideas and helped shape my work throughout my academic career. It has been a pleasure working with everyone.

My family has offered undeniable support throughout my whole life, and this would not have been possible without their unconditional love. I thank Tiffany Yu, my college best friend, who has been there for me ever since. I am grateful for Andrey Bzikadze, my significant other, for his companionship and his unwavering belief in me.

Chapter 1, in full, is a reprint of the material as it appears in Maksimov, M. O.*, Wu, C.*, Ashbrook, D. G., Villani, F., Colonna, V., Mousavi, N., Ma, N., Lu, L., Pritchard, J. K., Goren, A., Williams, R. W., Palmer, A. A., Gymrek, M. (2023). A novel quantitative trait locus implicates *Msh3* in the propensity for genome-wide short tandem repeat expansions in mice.

Genome Research, 33(5): 689-702. The dissertation author was one of the two lead investigators and authors of this paper.

Chapter 2, in full, is currently being prepared for submission for publication of the material by Wu, C., Xu, T., Munro, D., Mohammadi, P., Palmer, A. A., Goren, A, Gymrek, M. The dissertation author was the primary researcher and author of this paper.

Chapter 3, contains unpublished material by Wu, C., Shleizer-Burko, S., Goren, A, Gymrek, M. The dissertation author was the primary author of this chapter.

VITA

2017 Bachelor of Science in Biology with Specialization in Bioinformatics, University of California San Diego

2020 Master of Science in Computer Science, University of California San Diego

2023 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Maksimov, M. O.*, **Wu, C.***, Ashbrook, D. G., Villani, F., Colonna, V., Mousavi, N., Ma, N., Lu, L., Pritchard, J. K., Goren, A., Williams, R. W., Palmer, A. A., Gymrek, M. (2023). A novel quantitative trait locus implicates *Msh3* in the propensity for genome-wide short tandem repeat expansions in mice. *Genome Research,* 33(5): 689-702 (*these authors contributed equally)

Zheng, A., Lamkin, M., Zhao, H., **Wu, C.**, Su, H., & Gymrek, M. (2021). Deep neural networks identify sequence context features predictive of transcription factor binding. *Nature machine intelligence*, 3(2), 172-180.

Pope, W. H., Bowman, C. A., Russell, D. A., Jacobs-Sera, D., Asai, D. J., Cresawn, S. G., Jacobs, W. R., Hendrix, R. W., Lawrence, J. G., Hatfull, G. F.; **Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science**, Phage Hunters Integrating Research and Education, Mycobacterial Genetics Course. (2015). Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity, *eLife*. 28;4:e06416.

ABSTRACT OF THE DISSERTATION

Interrogating the effect of variation in *trans*-acting factors

by

Cynthia Wu

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Melissa Gymrek, Chair
Professor Alon Goren, Co-Chair

Mutations in *trans-acting factors* such as transcription factors, chromatin regulators, and DNA repair genes may cause widespread transcriptomic changes or altered genome-wide mutation patterns leading to a variety of phenotypes with varying impact on human health. Functional consequences of these mutations are difficult to systematically evaluate on a large scale due to the many challenges of *trans* studies. In this dissertation, I focus on two types of *trans*-acting factors: DNA repair genes, which control genome-wide mutation signatures, and transcriptional regulators, which impact genome-wide expression patterns. First, I present a study in which we used an unbiased genome-wide scan for regulators of repeat expansion propensity to identify the mismatch repair protein MSH3 as a strong *trans*-acting factor affecting germline mutation patterns in recombinant inbred mice. We found that inherited variants in and near *Msh3* result in variable mutation patterns that are most pronounced at long tetranucleotide repeats. Importantly, we also demonstrate a potential evolutionary tradeoff in which elevated *Msh3* leads to increased repeat expansions whereas *Msh3* deficiency results in a higher rate of short insertions and deletions. Second, I introduce a novel *trans*-eQTL detection method, xQTL, which is based on a biologically plausible mixture model of target gene effects sizes and results in increased power compared to traditional *trans*-eQTL analysis approaches. We applied xQTL to whole brain RNA-sequencing data from a cohort of outbred rats and identified 45 *trans*-eQTL candidates. For example, we identified a strong candidate *trans*-eQTL locus overlapping *Neurod4*, a key neuronal transcriptional factor, which xQTL estimate to regulate thousands of target genes. Importantly, this study also highlights key technical considerations regarding treatment of technical covariates when performing *trans*-eQTL detection. Last, I introduce scBE-seq which combines a pooled, high-precision genome editing strategy with single-cell sequencing assays to simultaneously interrogate the effects of hundreds of variants affecting *trans*-acting factors. Overall, these works furthered our understanding of the molecular effects of genetic variation on *trans*-acting factors and extended our toolkit for systematically studying their potential impact on complex phenotypes.

INTRODUCTION

A central question in genetics is how genetic variation can impact different traits, ranging from molecular phenotypes to disease status. Genetic variants may affect molecular phenotypes, such as gene expression, in either *cis* or *trans* (Farrall 2004; Signor and Nuzhdin 2018). In *cis* effects, a genetic variant influences a nearby region of the genome. For example, a variant nearby a gene may alter expression of that gene. On the other hand, *trans* effects of a variant impact distal regions of the genome, either on the same or different chromosomes. For example, a mutation in a transcription factor may result in gene expression changes in one or more of its target genes. In this work, we focus on the effects of mutations in *trans*.

Mutations in *trans*-acting factors have the potential to have widespread effects. There are various types of *trans*-acting factors such as transcription factors (TFs), chromatin regulators (CRs), splicing factors (SFs), and DNA repair genes. Mutations disrupting these factors, or expression levels of these factors can lead to global transcriptomic (TFs) or epigenomic changes (CRs), splicing variation (SFs), or altered mutation patterns (DNA repair genes) (Lee and Young 2013). For example, a single transcription factor may bind to thousands of genomic loci (Wang et al. 2012), potentially regulating many genes nearby these binding sites (Spitz and Furlong 2012; Lambert et al. 2018). Similarly, mutations in DNA repair genes can lead to accumulation of a large number of mutations throughout the genome ("mutator phenotypes") (Lipkin et al. 2000; Loeb et al. 2003; Pinto et al. 2013; Tome et al. 2013; Usdin et al. 2015).

Importantly, mutations in *trans*-acting factors have been implicated in human disease. They can lead to disorders with a wide range of severities from rare Mendelian disease to common complex traits and cancers. For example, mutations in transcription factor HNF4A cause maturity-onset diabetes of the young (Barrio et al. 2002) and mutations in the chromatin regulator EP300 are associated with the Mendelian disorder Rubinstein-Taybi syndrome (Roelfsema et al. 2005; Hamilton et al. 2016). Further, mutations in DNA repair genes MSH2 and other mismatch proteins have been associated with Lynch syndrome (Lynch et al. 2015), and mutations in ERCC2 have been linked to Trichothiodystrophy and Xeroderma pigmentosum (Cleaver et al. 1999; Cleaver et al. 2009). *Trans* effects have also been associated with complex traits. For example, a *cis*-eQTL for transcription factor KLF14 have been shown to have widespread *trans* effects in adipose and act as a risk modifier for cardiometabolic traits (Small et al. 2011). Overall, it has been estimated that approximately 60-70% of the heritability of gene expression is due to *trans*, rather than *cis*, variation (Grundberg et al. 2012).

Despite clear evidence of the importance of *trans* effects, identifying *trans*-acting mutations, particularly in population genetics studies, remain technically challenging and as a result, most such studies have focused on *cis* effects. For example, *trans*-eQTL studies are typically underpowered due to the large number of possible variant-gene associations to test which results in the multiple hypothesis burden whereas in *cis*-eQTL studies, the search space is limited to nearby genes (Huang et al. 2018). Importantly, *trans* effects are also generally weaker than *cis* effects (Pierce et al. 2014; Shan et al. 2019). Other factors such as known and unknown sources of technical variation, lack of comparable cohorts and tissues have resulted in false positive *trans*-eQTL calls and low reproducibility in human datasets (Gibson 2008; Innocenti et al. 2011; Stegle et al. 2012; Consortium et al. 2017).

Previous studies have focused on detecting *trans*-eQTLs with various methods. A study used the traditional *trans*-eQTL detection method by testing putative gene-by-variant pairs in yeast segregants and found *trans*-eQTL clustering at 102 hotspot loci (Albert et al. 2018). However, pairwise testing has less power due to the number of tests performed. Instead, methods have looked at various ways to evaluate one variant's impact on all genes. Another study tested for association between variants and aggregate representations of expression of gene sets based on various co-expression methods (Kolberg et al. 2020). This approach identified multiple *trans*-eQTLs in blood cell types for humans that were replicated in other studies. Yet, results were highly dependent on which co-expression method was chosen for analysis. A different study (Brynedal et al. 2017) leveraged cross-phenotype meta-analysis (CPMA) (Cotsapas et al. 2011) to identify global effects of a single variant by testing if the association statistics from all genes for the variant departs from the expected distribution under the null hypothesis of no *trans* effects. One limitation of CPMA is that this approach is best suited for detecting *trans*-eQTLs influencing many genes and has low power to detect *trans*-eQTLs with a small number of target genes which might be a more biologically plausible scenario. Additionally, these methods do not explicitly handle technical covariates that can confound true *trans*-eQTL signals.

Power for detecting *trans*-effects can be improved by increasing sample sizes and reducing experimental noise (Yao et al. 2020). One of the largest human expression datasets to date is from the Genotype–Tissue Expression (GTEx) consortium which includes a large cohort of ~800 donors with RNA sequencing from comparable tissues and whole genome sequencing, enabling characterization of genetic effects underlying human traits and diseases (Consortium 2020). However, many human studies with larger sample sizes are subject to environmental factors that are not possible to control, making it difficult to identify true genetic effects. Further, due to ethical reasons, samples for many tissues can only be collected postmortem.

However, it has been shown that different tissues have different responses and processes that occur over time elapsed since death (Ferreira et al. 2018). Altogether, failure to control for appropriate covariates, which are often unknown, can obfuscate true *trans*-eQTL signals.

Many issues present in human datasets can be addressed with model organisms. Animal models have allowed researchers to manipulate environmental factors to understand how they contribute to behaviors, traits, and diseases (Phillips and Roth 2019). Specifically, heterogeneity can be reduced, genetic variation can be constrained to common alleles, and experimental subjects can be exposed to certain conditions or substances that are not ethical to study with humans (Mukherjee et al. 2022). Mouse and rat are both ideal animal models due to the many similarities to humans in terms of anatomy and physiology (Vandamme 2014). Approximately 95% of genes are shared among the three species. Rodents are also relatively easy and cost effective to maintain and have short gestation periods and many offspring (Bryda 2013). Various model organisms offer their own advantages and disadvantages to studying human phenotypes.

Here, we focus on two rodent cohorts which have unique advantages for the specific phenotypes we are studying. The BXD mouse cohort consist of strains that have been inbred between the C57BL/6J (B) and DBA/2J (D) strains (Ashbrook et al. 2022). This cohort is ideal for studying regulators of mutation processes because they can be used to study mutations that have accumulated over many generations under controlled settings. Heterogeneous stock (HS) rats are outbred and have genomes made of a patchwork of eight founder haplotypes (Munro et al. 2022). There are abundant datasets for HS rats and their similar genetic structure to humans are suitable for studying transcriptional regulators impacting genome-wide expression patterns.

In this dissertation, I present three chapters that aim to further our understanding of *trans*-acting factors and their significance to specific phenotypes. I focus on two types of *trans*-acting factors: DNA repair genes, which control genome-wide mutation signatures, and transcriptional regulators, which impact genome-wide expression patterns. In chapter 1, my colleagues and I performed an unbiased genome-wide scan for regulators of repeat expansion propensity and identified the mismatch repair protein MSH3 as a strong *trans*-acting factor affecting germline mutation patterns in recombinant inbred mice. We found that inherited variants in and near *Msh3* result in variable mutation patterns that are most pronounced at long tetranucleotide repeats. Importantly, we also demonstrate a potential evolutionary tradeoff in which elevated *Msh3* leads to increased repeat expansions whereas *Msh3* deficiency results in a higher rate of short insertions and deletions.

Chapter 2 introduces xQTL, a novel *trans*-eQTL detection method that improves statistical power over traditional methods that test gene-by-variant pairs separately by jointly modeling effects of an individual variant across all genes. xQTL is based on a biologically plausible mixture model of target gene effect sizes. We applied xQTL on a whole brain RNA-sequencing dataset from a cohort of outbred rats and identified 45 *trans*-eQTL candidates. For example, we identified a strong candidate *trans*-eQTL locus overlapping *Neurod4*, a key neuronal transcriptional factor, which xQTL estimate to regulate thousands of target genes. Importantly, this study also highlights key technical considerations regarding treatment of technical covariates when performing *trans*-eQTL detection.

Chapter 3 presents our efforts to develop scBE-seq (single cell base editing sequencing) which combines a pooled, high-precision genome editing strategy with single-cell sequencing assays to simultaneously interrogate the effects of hundreds of variants. Specific mutations can have severe health consequences, while other mutations in the same gene can have little to no impact. The assay aims to advance our understanding of the regulatory consequences of genetic variation in *trans*-acting regulators and provides a complementary experimental approach in addition to the computational approaches in Chapter 1 and 2. Furthermore, scBE-seq allows us to validate potential candidate mutations in *trans*-acting factors identified in Chapter 2 in cell types of interest. Here, I present my initial developments of scBE-seq and discuss its ongoing progress.

Overall, these three chapters aim to further our knowledge of the effect and our ability to detect and study genetic variation in *trans*-acting factors. Chapter 1 focuses on mutations in DNA repair genes, associated with genome-wide mutation signatures. Chapter 2 studies mutations in transcriptional regulators impacting genome-wide expression patterns. Chapter 3 provides an experimental assay to interrogate mutations of interest in *trans*-acting factors which can be used to validate findings of Chapter 1 and 2.

CHAPTER 1

A novel quantitative trait locus implicates Msh3 in the propensity for genome-wide short tandem repeat expansions in mice

1.1 Abstract

Short tandem repeats (STRs) are a class of rapidly mutating genetic elements typically characterized by repeated units of 1–6 bp. We leveraged whole-genome sequencing data for 152 recombinant inbred (RI) strains from the BXD family of mice to map loci that modulate genome-wide patterns of new mutations arising during parent-to-offspring transmission at STRs. We defined quantitative phenotypes describing the numbers and types of germline STR mutations in each strain and performed quantitative trait locus (QTL) analyses for each of these phenotypes. We identified a locus on Chromosome 13 at which strains inheriting the C57BL/6J (B) haplotype have a higher rate of STR expansions than those inheriting the DBA/2J (D) haplotype. The strongest candidate gene in this locus is *Msh3*, a known modifier of STR stability in cancer and at pathogenic repeat expansions in mice and humans, as well as a current drug target against Huntington's disease. The D haplotype at this locus harbors a cluster of variants near the 5′ end of *Msh3*, including multiple missense variants near the DNA mismatch recognition domain. In contrast, the B haplotype contains a unique retrotransposon insertion. The rate of expansion covaries positively with *Msh3* expression—with higher expression from the B haplotype. Finally, detailed analysis of mutation patterns showed that strains carrying the B allele have higher expansion rates, but slightly lower overall total mutation rates, compared with those with the D allele, particularly at tetranucleotide repeats. Our results suggest an important role for inherited variants in *Msh3* in modulating genome-wide patterns of germline mutations at STRs.

1.2 Introduction

Studies of germline and somatic mutations have shown considerable variation across individuals and species in both the rate and patterns by which mutations occur (Lynch et al. 2016). In some cases, this variation may be controlled by heritable factors influencing the function or expression of proteins involved in maintaining genome integrity. Indeed, genetic variants have been identified that disrupt DNA repair proteins (Taylor et al. 1997; Li 2008) and lead to "mutator" phenotypes in which affected individuals or cells accumulate specific types of mutations at a faster rate. Although some of these phenotypes are highly deleterious, such as in cancer, common genetic variation can also result in more moderate mutator phenotypes that are only identified upon molecular interrogation (Sasani et al. 2022). Identifying genetic factors controlling this variation can provide insight into mutation processes and DNA repair mechanisms.

Short tandem repeats (STRs), typically consisting of repeated sequence motifs of 1–6 bp, show rapid mutation rates that are orders of magnitude greater than those for single-nucleotide variants (SNVs) (Sun et al. 2012). STR mutations typically result in expansions or contractions of one or more copies of the repeat unit. Expansion mutations are well known to cause a variety of disorders in humans, including Huntington's disease, hereditary ataxias, and myotonic dystrophy (Hannan 2018). Further, we and others have recently implicated both small and large expansions and contractions at STRs in autism spectrum disorder (Trost et al. 2020; Mitra et al. 2021). Finally, frequent somatic mutations at STRs, referred to as microsatellite instability (MSI), are a hallmark of certain cancer types (Vilar and Gruber 2010).

A large number of disease-focused studies have implicated proteins involved in mismatch repair (MMR) in regulating STR stability. For example, Lynch syndrome, which results in a predisposition to colorectal and other cancer types characterized by MSI, can be caused by mutations that disrupt a variety of MMR proteins (Lynch et al. 2015). On the other hand, multiple MMR proteins (including MSH2, MSH3, MLH1, and MLH3) have been shown to be required for somatic expansions of CAG repeats in mice (Manley et al. 1999; López Castel et al. 2010; Pinto et al. 2013). Further, genome-wide association studies (GWASs) for the age of onset and progression of Huntington's disease have identified mutations in MLH1 (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2015) and MSH3 (Moss et al. 2017) that lead to increased somatic instability of the pathogenic trinucleotide expansion at HTT, and MSH3 is a current drug target for Huntington's disease (Kingwell 2021). Taken together, these studies suggest a critical role of inherited variation in MMR genes in regulating patterns of STR mutation.

The majority of studies of STR mutator phenotypes to date have focused on somatic repeat instability. However, studies of de novo STR and other mutation types have also shown considerable variation in germline mutation rates across individuals (Turner et al. 2017; Mitra et al. 2021). Although this variation is also potentially genetically controlled, this phenomenon is difficult to study in humans. Germline mutation rates are strongly confounded by parental age (Kong et al. 2012), and mutation spectra may be influenced by environmental exposures (Nik-Zainal et al. 2015). Further, observed mutation patterns in children result from a mixture of mutation processes in the maternal and paternal germline. Thus, the relevant genetic variation controlling germline mutations could be harbored by either of the parents and is challenging to study in a typical GWAS setting.

Inbred mouse strains offer a unique opportunity to determine regulators of mutation processes because they can be used to study mutations that have accumulated over many generations under controlled settings. Further, within each strain, offspring and both parents share essentially identical genomes, and thus, offspring and parental genotypes do not need to be considered separately. Here we focused on the BXD family (Ashbrook et al. 2021), which consists of strains that were generated by serial inbreeding of progeny of crosses between inbred C57BL/6J (B) and DBA/2J (D) strains. Strains were generated in multiple rounds ("epochs") by different groups spanning several decades (Ashbrook et al. 2021), during which STR and other mutations have accumulated in the resulting strains. We leveraged genome-wide STR genotypes generated from whole-genome sequencing (WGS) of the BXD family (Ashbrook et al. 2022) to determine the contribution of inherited genetic variation to the number and patterns of new STR mutations across the genome arising during parent-to-offspring transmission.

1.3 Results

1.3.1 Identifying new mutations in the BXD family

We previously built a reference set of 1,176,016 autosomal tandem repeats consisting of 1,154,738 STRs (repeat unit 2–6 bp) and 21,278 variable number tandem repeats (repeat unit 7+ bp) identified from the mm10 (C57BL/6J) reference assembly, and applied GangSTR (Mousavi et al. 2019) to genotype these STRs using WGS of 152 strains from the BXD cohort (Ashbrook et al. 2022). Homopolymer repeats (repeat unit 1 bp) were excluded as we could not obtain reliable genotypes for those loci in this cohort, which was not generated using PCR-free protocols. For simplicity, we refer below to all repeats analyzed as STRs, because the majority have repeat units <7 bp. We used these genotypes to identify new germline STR mutations by comparing the genotype at each strain to that expected based on the founder haplotype at that region. The majority of accumulated mutations likely arose over previous generations of inbreeding and are expected to be homozygous as the BXD strains have been inbred for up to

180 generations. Although heterozygous genotypes may represent true recent mutations, they were removed from downstream analysis because these are likely enriched for STR genotyping errors. In total, we identified 18,119 unique loci (18,053 STRs and 66 VNTRs) for which at least one BXD strain is homozygous for an STR length that does not match the expected founder genotype, indicating a candidate new mutation (Fig. 1.1A; Supplemental Datasets S1–S3). These mutations may occur at STRs for which both founders harbored the same allele or may occur at STRs that were already polymorphic in the founders. Mutations are scattered throughout the genome and do not cluster at any particular genomic location (Supplemental Fig. 1.6). Most mutations identified occur at tetranucleotide STRs, which are also most highly represented among successfully genotyped loci (Supplemental Fig. 1.7A). Dinucleotide STRs, which are uniquely abundant in many rodent genomes (Srivastava et al. 2019), are underrepresented in our data set as a consequence of filtering due to low genotyping quality.

We used SNP genotypes surrounding each STR to determine whether the mutation occurred on the parental B or D haplotypes, which enabled us to accurately determine the size of each mutation. We observed a slight excess of new mutations originating on B haplotypes (52.5%) (Supplemental Fig. 1.7B), consistent with an overall slight excess of B haplotypes within the family. Most mutations result in expansions or contractions of a single repeat unit compared with the founder, with a bias toward expansion mutations (Fig. 1.1B). Mutations of two or more repeat units are slightly more prevalent among dinucleotide and trinucleotide repeats than among tetranucleotide repeats (Supplemental Fig. 1.7C). Both trends are consistent with those seen in human de novo STR mutations (Sun et al. 2012; Mitra et al. 2021). Nearly all mutations identified result in expansion or contraction by at most five repeat units, although our pipeline is not optimized to identify larger expansions.

Observed STR mutations are consistent with the known history of generation of the BXD family. The BXD strains are divided into epochs, corresponding to various rounds of strain generation occurring from 1970 to 2014 (Ashbrook et al. 2021). Assuming, for simplicity, a constant mutation rate per generation, the number of candidate STR mutations is expected to increase with the number of generations of inbreeding (Fig. 1.1C). Although 58% of new mutations identified are private to a single strain, the remainder are found in two or more strains (Supplemental Fig. 1.8). Principal components analysis (PCA) based on genotypes at STRs for which we observe new mutations clearly separates strains by epoch (Fig. 1.1D), indicating that some STR mutations are epoch specific and arose in parental stocks ancestral to each successive epoch.

1.3.2 Mapping quantitative trait loci for STR mutation phenotypes

We wondered whether observed differences in the number and size of mutations across strains could be driven by genetic variation affecting DNA repair or other pathways. To this end, we defined several quantitative phenotypes to summarize STR mutation patterns in each strain. We focused on three basic characteristics. *Mutation count* was computed as the fraction of genotyped STRs with a new mutation in each strain. Notably, this does not truly represent a germline de novo mutation rate, because observed mutations are homozygous and therefore must have occurred in ancestors to present-day individuals used for sequencing. *Mutation size* was calculated as the average change in repeat unit count, computed separately for expanded versus contracted mutations in each strain. *Expansion propensity* was calculated as the fraction of new mutations in each strain for which the new allele is longer than the founder allele (the same phenotype could be defined for *contraction propensity*, but this is redundant as it is simply 1 – expansion propensity). For all phenotypes, we filtered new mutations seen in more than 10

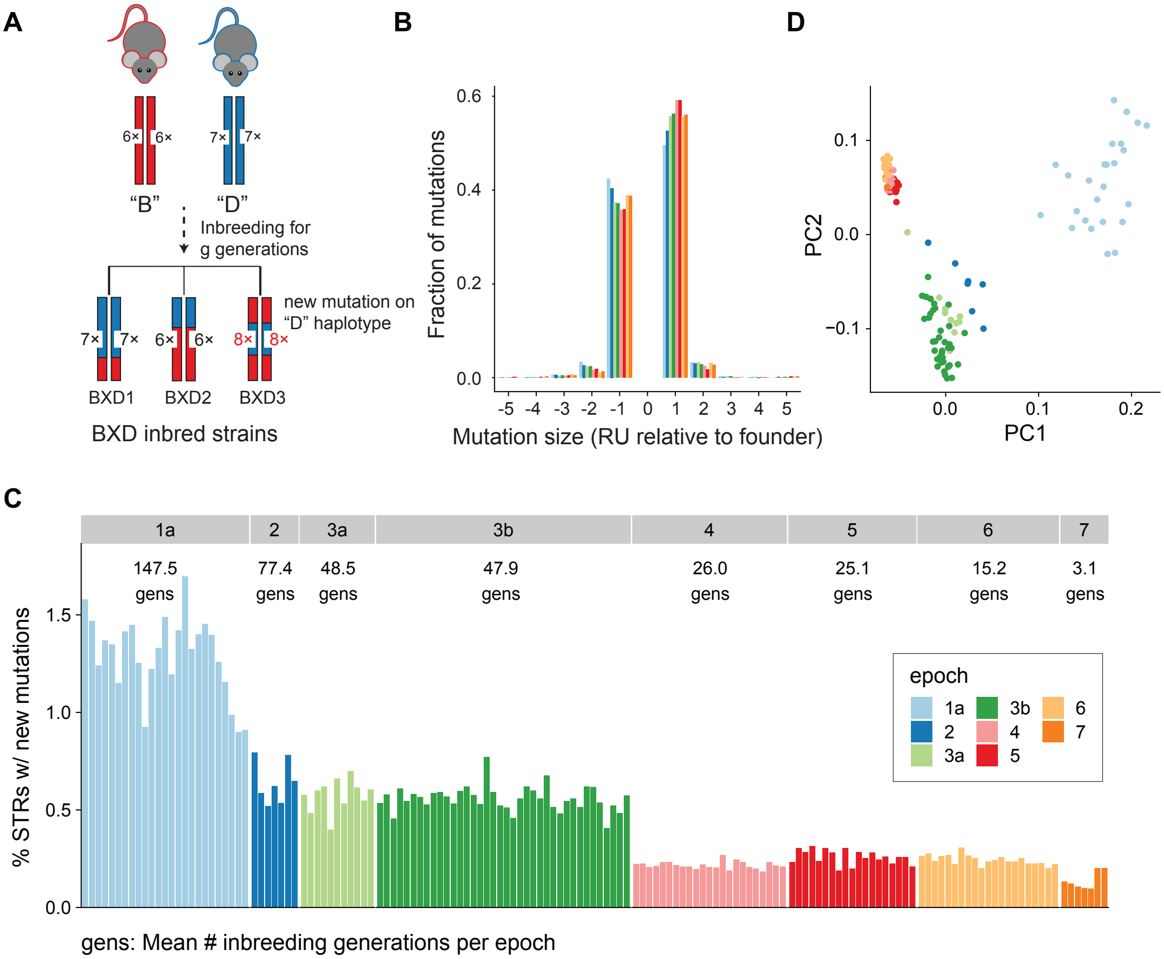
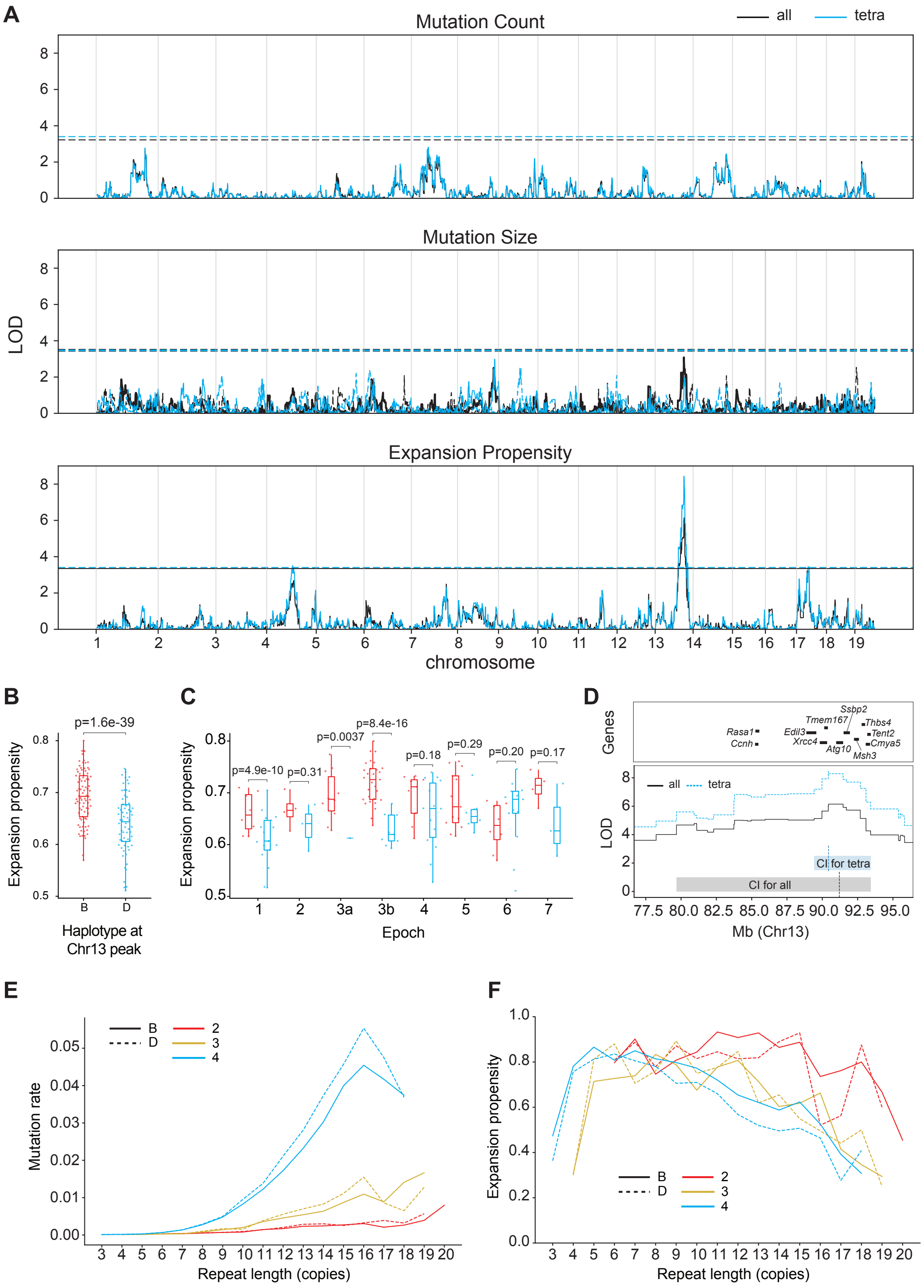


Figure 1.1 Characterizing new mutations in the BXD family. (A) Schematic of new mutation discovery. Each strain's genome is a homozygous patchwork of segments derived from multiple generations of inbreeding of the descendants of the founders, C57BL/6J (B; red) and DBA/2J (D; blue). A full description of the breeding history for each epoch is described in Supplemental Figure S1 of Ashbrook et al. (2021). Our STR mutation discovery pipeline considers a fixed set of STRs discovered in the mm10 reference genome (in the example shown, B has six copies and D has seven copies of the repeat for a particular STR). We identify new mutations as STRs with repeat lengths differing from the length of the founder inferred at that genome segment. In the example, strain BXD3 has a mutation to eight copies that occurred on a haplotype inherited from the D founder. **(B)** Distribution of mutation sizes for each BXD epoch. The x-axis shows mutation sizes in terms of the difference in number of repeat units (RUs) from the founder allele. Positive sizes indicate expansions, and negative sizes indicate contractions. Distributions are calculated separately for strains belonging to different epochs, indicated by bar color. Mutations range in size from –16 to +9 RUs, but plots are restricted to ±5 because 99.9% (52,784/52,812) of observed mutations fall in this range. **(C)** Percentage of genotyped STRs with a new mutation for each strain. New mutations refer to any STR for which the observed allele does not match the expected founder allele. The average number of generations of inbreeding for strains is annotated for each epoch. Strains are sorted by decreasing numbers of inbreeding generations within each epoch. **(D)** Principal component analysis (PCA) of new mutations. PCA was performed on a binary matrix indicating whether each strain does or does not carry the new allele at each STR. The first two principal components separate strains by epoch, indicating combinations of new mutations are shared among strains in each group. For B–D, colors denote BXD epochs, as annotated in panel C.

strains, because those have likely been segregating within the BXD family on a variety of genetic backgrounds that differ from that of the individual in which the mutation initially arose. These common mutations may also represent cases in which the founder was incorrectly genotyped, leading to false-positive mutation calls. Because of their high mutation rates, recurrent mutations are expected, and so we did not restrict our analysis to mutations seen only once in our cohort. We further restricted analysis to strains with at least 10 observed mutations because mutation phenotype values are unreliable when computed over a small number of mutations.

We performed genome-wide QTL mapping separately for each of these mutation phenotypes using R/QTL2 (Broman et al. 2019) and a set of 7,101 LD-pruned SNPs (Fig. 1.2). To account for population structure, R/QTL2 uses a linear mixed model with a kinship matrix generated using the leave-one-chromosome-out (LOCO) approach. The number of generations of inbreeding for each strain was used as a covariate. We determined genome-wide significance thresholds based on permutation analysis. QTL analysis did not identify any genome-wide significant loci for mutation size or mutation count. However, we identified a strong signal on Chr 13 (max logarithm of the odds [LOD] = 6.1) associated with expansion propensity. Strains with the B haplotype at this locus tend to have higher expansion propensity than those with the D haplotype (Fig. 1.2B,C). This trend is consistently observed when considering mutations in either genic or intergenic regions (Supplemental Fig. 1.9). The QTL is centered at 91.2 Mb with a 1.5-LOD support interval from 79.7 to 93.4 Mb, a region that encompasses several dozen genes (Fig. 1.2D; Supplemental Table S1). Two additional suggestive peaks were identified for expansion propensity on Chr 4 and Chr 17 (Supplemental Fig. 1.10). To investigate whether the strongest expansion propensity signal might be driven by specific types of STRs, we

Figure 1.2 Discovery of QTLs for STR mutation phenotypes. (A) QTL mapping results. Panels show results for mutation count (top), mutation size (middle), and expansion propensity (bottom). The x-axis shows the genomic location, and the y-axis shows the LOD score of each SNP. For mutation size, solid traces and dashed traces represent contraction and expansion mutations, respectively. For each panel, black indicates the phenotype based on all STRs; blue, the phenotype based on tetranucleotide STRs only. Dashed horizontal lines show genome-wide significance thresholds based on permutation analyses. **(B,C)** Increased expansion propensity is associated with the B haplotype at the Chr 13 QTL. Each point represents one strain. We used SNP haplotype blocks to assign each strain as harboring either the B (red) or D (blue) haplotype at this locus. The y-axis denotes expansion propensity. Panel B shows the trend across all BXD strains, and panel C shows the trend separately for each epoch. Horizontal lines show median values; boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to the minimum and maximum data points in each group. For panels B and C, annotated P-values are based on a two-sided z-proportion test. **(D)** Genes located in or near the QTL peak. The y-axis shows the QTL signal (LOD score) for expansion propensity at Chr 13. Black line indicates all STRs; blue line, tetranucleotide STRs. Shaded boxes indicate the 1.5-LOD confidence interval for all STRs (gray box) and tetranucleotides (light blue box). Horizontal bars denote a subset of genes near the center of the QTL peak. A full list of genes in this region is given in Supplemental Table S1. **(E)** Repeat length versus relative mutation rate. The x-axis gives the repeat length of each STR based on the parent haplotype at each locus in each strain. The y-axis gives the relative mutation rate of STRs in each bin, computed as the number of mutations divided by the total number of nonmissing genotype calls falling in each bin. **(F)** Repeat length versus expansion propensity. The x-axis is the same as in E. The y-axis gives the proportion of mutations observed in each bin that are expansions. For E and F, red indicates dinucleotides; gold, trinucleotides; and blue, tetranucleotides. Dashed lines indicate D haplotype; solid lines, B haplotype at the Chr 13 QTL locus.


A
Mutation Count
all
tetra
Mutation Size
LOD
Expansion Propensity
0
2
4
6
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
chromosome
B
p=1.6e-39
Expansion propensity
0.5
0.6
0.7
0.8
B
D
Haplotype at Chr13 peak
C
p=4.9e-10
p=0.31
p=0.0037
p=8.4e-16
p=0.18
p=0.29
p=0.20
p=0.17
Expansion propensity
1 2 3a 3b 4 5 6 7
Epoch
D
Genes
Rasa1
Ccnh
Edil3
Xrcc4
Tmem167
Ssbp2
Atg10
Msh3
Thbs4
Tent2
Cmya5
all
tetra
LOD
CI for tetra
CI for all
77.5 80.0 82.5 85.0 87.5 90.0 92.5 95.0
Mb (Chr13)
E
B
D
2
3
4
Mutation rate
0.00
0.01
0.02
0.03
0.04
0.05
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Repeat length (copies)
F
Expansion propensity
0.0
0.2
0.4
0.6
0.8
1.0
B
D
2
3
4
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Repeat length (copies)

repeated QTL mapping separately for each repeat unit length. The signal is strongest by far for tetranucleotide STRs (max LOD = 8.4; 1.5-LOD support interval, 89.4–93.4) (Supplemental Fig. 1.11), which are the most abundant STR type in our data set. Notably, all but tetranucleotide STRs have overall low mutation counts, resulting in unreliable estimates of expansion propensity for those categories (Supplemental Fig. 1.12). When tested individually, both di- and tetranucleotides showed at least nominally significant signals (two sided z-proportion test P = 0.038 and P = 3.7×10^{38}, respectively), but trinucleotides did not (P = 0.95).

To test whether the Chr 13 signal is influenced by our choice of filtering parameters, we repeated QTL mapping using a range of thresholds for the minimum number of mutations observed per strain and the maximum number of strains in which each new mutation was identified (Supplemental Fig. 1.11). Overall, the signal is robust to these filters and increases as we restrict analysis to successively rarer mutations. However, the signal is weaker when considering only private variants, which could be due to a combination of reduced power from lower mutation counts and enrichment of genotyping errors at private mutations. We additionally tested whether the observed signal replicates across BXD epochs, which were generated at separate times and locations and could potentially have different environmental exposures or epoch-specific variants driving mutator phenotypes. The Chr 13 signal is strongest in epoch 3b, which has the most strains and therefore is the best powered. Additionally, epochs 1 and 3a show significant signals when tested individually (Fig. 1.2C), and the signal is strongest when including all epochs (Supplemental Fig. 1.13). Further, the direction of effect is consistent across most epochs, with the exception of later epochs for which a smaller number of mutations have accumulated (Fig. 1.2C). Thus, we concluded the causal variant is segregating across the entire BXD family, and the QTL is not due to an epoch-specific mutation or environmental phenomenon.

We then investigated genome-wide STR mutation patterns and whether these are influenced by the haplotype at the Chr 13 locus. For all repeat unit lengths (2–4 bp), relative mutation rate increases as a function of the total length of the repeat (Pearson $r = 0.93$, 0.94, 0.93 for di-, tri-, and tetranucleotide loci, respectively, with $P < 10^6$ in all cases) (Fig. 1.2E; Supplemental Fig. 1.14), consistent with many previous observations of STR mutation patterns (Payseur et al. 2011; Sun et al. 2012). Tetranucleotides showed the highest overall mutation rates, followed by trinucleotides and dinucleotides. However, because many highly polymorphic dinucleotides were excluded from analysis owing to low-quality genotypes (see Methods), observed relative mutation rates are likely underestimated for those loci. Although we did not observe a genome-wide significant association between the Chr 13 signal and mutation count (Fig. 1.2A), we observed that longer repeats (parent repeat length, ~>30 bp) tended to show higher mutation rates in strains carrying D haplotypes for the QTL. We found that this trend of higher mutation rates for the D alleles remains when considering only mutations arising on either B or D local haplotype backgrounds (Supplemental Fig. 1.14), and therefore, it is not biased by the fact that the B haplotype matches the mm10 reference genome. Stratifying by repeat unit sequence showed that AGAT repeats have the highest mutation rates across both groups. AGAT, AAAC, AAAT, and ACAT repeats have significantly higher mutation rates in strains with the D haplotype (two-sided z-proportion test $P < 0.05$) (Supplemental Fig. 1.15), with trends in the same direction for the majority of other repeat unit sequences.

We further examined expansion propensity as a function of repeat length. The rate of expansion is negatively associated with total repeat length (Pearson r = –0.60, –0.47, –0.66 and P = 0.019, 0.054, 0.0052 for di-, tri-, and tetranucleotides) (Fig. 1.2F), indicating longer repeats have a higher tendency to contract relative to shorter repeats. Consistent with the association signal for expansion propensity described above, we found mutations at tetranucleotide STRs in strains with the B haplotype at the Chr 13 QTL have a higher probability to be expansions across a broad range of repeat lengths (Fig. 1.2F; Supplemental Fig. 1.14). We also observed a suggestive signal in the expansion propensity QTL region for contraction size (Fig. 1.2A) and found that contraction mutations tend to be larger for strains with the B Chr 13 QTL haplotype, whereas the size of expansion mutations is similar between groups (Supplemental Fig. 1.14). Stratifying by repeat unit sequence shows that AGAT and AAAT repeats show the most significant differences in repeat expansion propensity between strains with the B versus D haplotype (two-sided z-proportion test $P < 0.05$), but suggestive trends in the same direction are observed for most other repeat units (Supplemental Fig. 1.15).

Finally, we investigated whether the observed expansion propensity signal might be driven primarily by mutations arising in either the maternal or paternal germline by comparing the patterns of STR mutations on autosomes versus the two sex chromosomes. Intuitively, if the signal is driven primarily by mutations in the female germline, we would expect to see no impact on Chr Y for which all mutations are paternal germline derived, but a stronger signal on Chr X for which two-thirds of mutations are expected to be maternal. In contrast, if the signal is driven by the paternal germline, we would expect to see the strongest signal for Chr Y mutations and the weakest signal for Chr X. A total of 1,228 mutations at 666 unique STRs were identified on Chr X and Chr Y. For all scenarios tested, expansion propensity was significantly higher for strains with the B versus D haplotype of the Chr 13 QTL, irrespective of chromosome (Supplemental Fig. 1.16). Although the magnitude of this trend is strongest for Chr Y, the difference between B and D is not statistically significant (two-sided z-proportion test $P > 0.05$) for both sex chromosomes. However, this analysis may be underpowered owing to the smaller number of mutations on sex chromosomes. Overall, our results are suggestive of a paternal origin effect, but other parent of origin scenarios cannot be ruled out.

1.3.3 Analysis of candidate variants disrupting protein-coding genes

We next sought to characterize the QTL on Chr 13 for expansion propensity identified above. We first searched for variants predicted to impact gene function that fall within the QTL 1.5-LOD support interval for the tetranucleotide signal. We identified 5,982 SNPs/indels and 214 STRs overlapping protein-coding genes. We additionally performed pangenome analysis to identify 3,698 large structural variants (SVs; 50 bp < SV < 10 kbp) (Supplemental Fig. 1.17). To reduce the search space, we removed rare variants (nonmajor-allele fraction < 0.15) and variants only weakly associated with the expansion propensity phenotype (model P-value > 5×10^{4}). We used the Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016) to annotate the predicted impact (modifier, low, moderate, or high) of the 5,250 variants that remained after filtering (Supplemental Tables S1–S3; Supplemental Fig. 1.18).

Based on previous studies of STR instability in cancer (Lynch et al. 2015) or modifiers of repeat expansion disorders (Wheeler and Dion 2021), we hypothesized that the observed STR mutator phenotype might be driven by variation in DNA repair genes. Of the genes in the QTL region, four are known to be involved in processes related to DNA repair: *Xrcc4* (nonhomologous end joining to repair double-strand breaks), *Ssbp2* (DNA damage response), *Atg10* (autophagy mediated effect) (Demirbağ -Sarikaya et al. 2021), and *Msh3* (involved in

MMR), which has been widely implicated in STR stability (Dragileva et al. 2009; Boland and Goel 2010; Tomé et al. 2013a).

Of DNA repair genes in this region, *Ssbp2* contains only variants marked as modifiers by VEP, which are unlikely to impact protein function directly, and *Xrcc4* contains multiple variants predicted to have low or moderate impact (Supplemental Table S2). *Atg10* has a more extensive variant profile with two moderate impact missense variants predicted as tolerated by SIFT (Sim et al. 2012), one low impact synonymous variant, and a multiallelic coding sequence insertion (Supplemental Table S2), with a common allele resulting in an in-frame insertion (rs230013535) and a rarer allele causing a frameshift. Closer inspection of the frameshift allele revealed that all four strains carrying the allele are heterozygous and have lower genotype quality scores than other strains at the locus, suggesting this allele is a variant calling artifact and unlikely to explain the QTL signal.

Msh3 contains the most variants with effects predicted by VEP, including one splice, four missense, and three synonymous mutations within protein-coding exons (although after normalizing for transcript length, *Xrcc4* contains slightly more variants per base pair) (Supplemental Table S1). Most of these are located within a variant-dense region in the 5′ end of the gene near the mismatch recognition domain (Supplemental Fig. S13; Supplemental Table S2; Fig. 1.3A) and have been previously shown to be associated with expansion propensity of CAG repeats in an *HTT trans*-gene (Tomé et al. 2013a). One of the missense variants (rs48140189) is predicted by SIFT to be deleterious within a truncated transcript but is tolerated within both canonical transcripts. In addition to impactful variants within protein-coding transcripts, we also identified three variants of interest mapping to a nonsense-mediated decay (NMD) transcript of *Msh3* (ENSMUST00000190393). One of these is a 387 bp insertion, corresponding to an IAPLTR2a retrotransposon (Thompson et al. 2016), in C57BL/6J compared with DBA/2J (Supplemental Fig. 1.19). The insertion spans nearly the entirety of exon 5 of the NMD transcript and falls in an intron between exons 4 and 5 of the canonical transcript (Supplemental Fig. 1.19B; Fig. 1.3B). The other two variants are adjacent to the IAPLTR2a insertion and could plausibly be driven by mapping artifacts in this region owing to the high density of nonuniquely mapped reads at retrotransposon elements. We further examined other SVs within each gene that passed the association and allele frequency criteria regardless of their impact predicted by VEP (Supplemental Table S4). Although *Atg10* and *Ssbp2* harbor several large (>50 bp) SVs with similarly large LOD scores, neither of these is predicted to overlap with any meaningful feature.

Finally, we identified several variants in proteins not involved in DNA repair that were predicted to have high impact (Supplemental Table S3). Two frameshift mutations were found in *Cmya5*, a gene primarily involved in muscle- and cardiac-related phenotypes (Lu et al. 2022) and thus unlikely to be related to an STR mutator phenotype. We additionally identified a stop loss mutation in *Zcchc9*, which encodes a zinc finger–containing protein that can bind DNA or RNA (Zhou et al. 2008). Although we cannot rule out the impact of this mutation, there is currently no known link between this gene and STR stability.

1.3.4 Expansion propensity QTL colocalizes with multiple cis-eQTLs

We next wondered if the QTL for expansion propensity might also be mediated through *cis*-regulatory variants affecting expression of genes in this region. To this end, we compiled 54 publicly available gene expression microarray data sets encompassing 30 tissues (Supplemental Table S5), with sample sizes ranging from 11–79 strains. Notably, these data sets were acquired using multiple microarray platforms, under different experimental

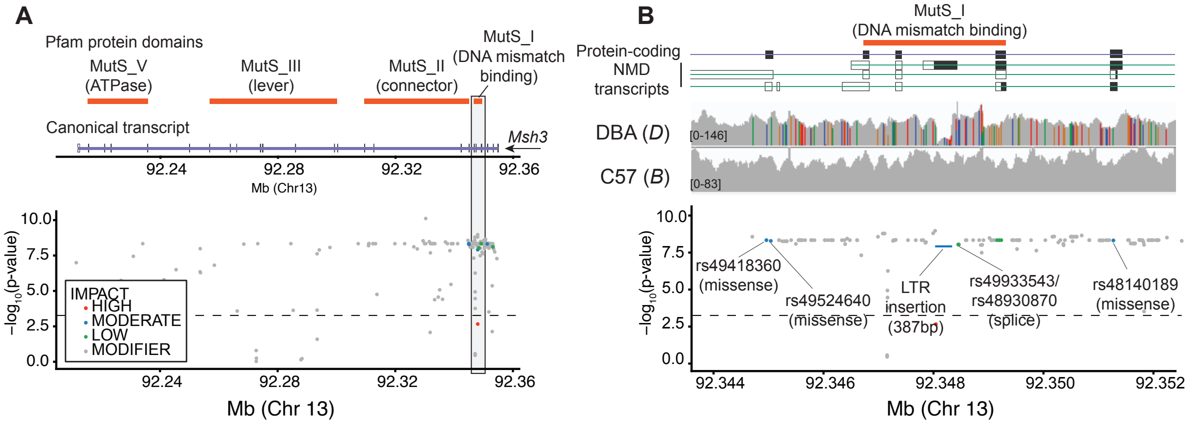


Figure 1.3 Variants predicted to impact *Msh3*. (A) Summary of variants overlapping *Msh3*. The *top* panel shows the canonical protein-coding transcript of *Msh3* (purple) and protein domains (orange rectangles) obtained from Pfam (Mistry et al. 2021). The *bottom* panel shows the location (mm10; x-axis) of variants and their association with the expansion propensity phenotype (–log10 P-values; y-axis). Variants are colored by their impact predicted by VEP: red indicates high; blue, moderate; green, low; gray, modifier). **(B)** Summary of variants in the variant-dense 5′ region of *Msh3*. *Top* and *bottom* panels are the same as in A. The *middle* panel shows a histogram of read coverage as visualized using the Integrative Genomics Viewer (Robinson et al. 2011). Colored bars denote the fraction of reads at each position with mismatches from the reference, which is based on C57BL/6J. Gray denotes matches to the reference. In both panels, rare variants are excluded (non-major-allele fraction < 0.15). The –log10 (P-value) threshold distinguishes variants associated with the expansion propensity phenotype (model P-value ≤ 5 × 10^4).

conditions and across a range of tissues. Overall, we find that *Ssbp2* is among the most highly expressed genes within the QTL region; *Msh3* has average expression; and *Atg10* and *Xrcc4* are expressed slightly below average (Supplemental Fig. 1.20). For downstream analyses, we restricted to 40 expression data sets with at least 30 strains. We found that the subset of BXD strains included in each of these expression data sets was in most cases sufficient to reproduce the expansion propensity QTL signal originally identified using all 152 strains, indicating the relevant causal variant(s) of interest are likely segregating in each of those subsets (Supplemental Fig. 1.21).

For each of these 40 expression data sets, we performed a separate expression QTL (eQTL) analysis for 25 protein-coding genes for which expression levels are available in at least half of the data sets (Supplemental Fig. 1.22). We considered only probes not overlapping SNPs for comparing gene expression levels and used the number of variants per probe as a covariate in eQTL mapping to avoid confounding the true variability with differences in probe hybridization efficiency. Notably, this excluded a large number of probes for *Msh3* because many overlap multiple SNPs in the highly variable 5′ end of the gene (Supplemental Fig. 1.23). We then ranked genes by the proportion of data sets in which the maximum eQTL LOD exceeded the permutation-based threshold for significance (Supplemental Fig. 1.24). We observed robust eQTL signals for *Ssbp2* and for *Atg10* in 29 and 18 data sets, respectively. We also found eQTL signals for *Xrcc4* and *Msh3*, albeit in a smaller number of data sets: six and four, respectively (Fig. 1.4A; Supplemental Fig. 1.25A). The eQTL for *Atg10* shows the most consistent colocalization with the QTL peak across data sets (Supplemental Fig. 1.25A). However, eQTLs for most genes in the region are strongly colocalized with the QTL (Fig. 1.4A;

Supplemental Fig. 1.26), making it difficult to prioritize a single causal gene based on the eQTL signal alone.

We further examined the eQTL signal at *Msh3*, given its previously reported role in STR stability (Campregher et al. 2012; Flower et al. 2019). In all tissues with a significant eQTL for *Msh3*, we observed a consistent direction of effect, with higher *Msh3* expression for strains carrying the B haplotype associated with increased expansion propensity (Fig. 1.4B; Supplemental Fig. 1.25B). Detailed analysis of the *Msh3* eQTL shows that the signal is strongest when considering probes and variants in the 5′ end, even after adjusting for hybridization efficiency owing to SNPs in this region (Methods; Supplemental Fig. 1.27). This result is consistent with previous studies in humans, in which increased *MSH3* expression driven by polymorphism in the 5′ end of the gene was associated with increased somatic instability at the trinucleotide repeat involved in Huntington's disease (Flower et al. 2019). Notably, *Dhfr*, which shares a promoter with *Msh3*, did not show a strong eQTL signal in the expression data sets tested (Supplemental Fig. 1.24).

Finally, we examined tissue-specific expression of each of the candidate DNA repair genes using the Bgee (Bastian et al. 2021) database (Supplemental Table S6). Although STR mutations here were assessed from spleen- and tail-derived DNA, we assume the majority result from transmission events along the germ lineage and, therefore, likely arose in tissues related to reproduction. *Msh3* is most highly expressed in reproductive (oocytes and spermatocytes) and zygotic tissues. On the other hand, *Atg10*, which is also near the QTL center, is most highly expressed in heart structures, which are unlikely to be relevant for germline mutations. *Ssbp2* is expressed in a variety of tissues, and Xrcc4 is expressed in spermatocytes and oocytes. However, variants overlapping Xrcc4 have lower LOD scores for association with expansion

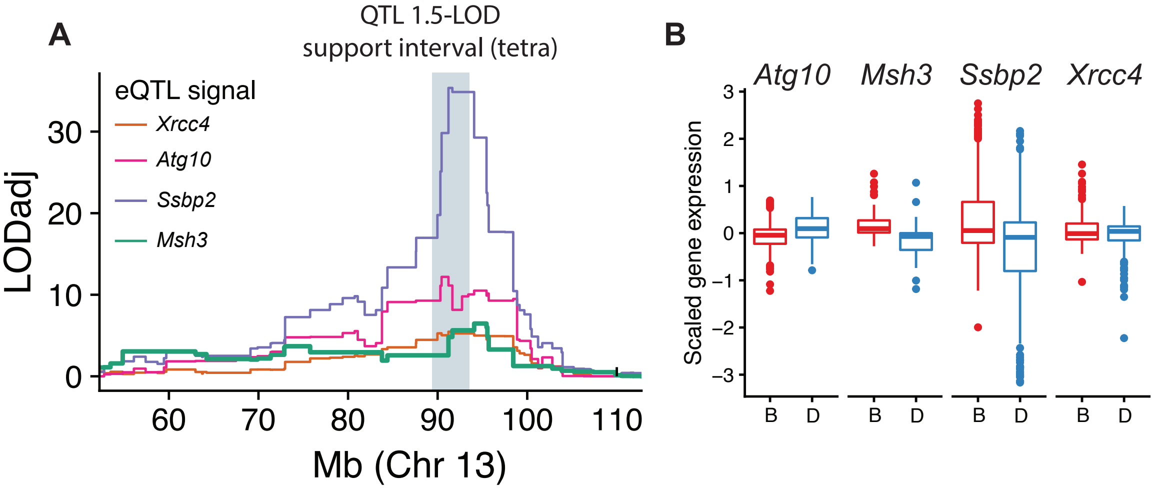


Figure 1.4 The Chr 13 expansion propensity QTL colocalizes with eQTLs for multiple DNA repair genes. (A) Colocalization of expansion propensity and eQTL signals. Colored traces denote eQTL LOD scores. Each line shows the expression data set with the strongest eQTL for that gene. eQTL LOD scores were adjusted for multiple hypothesis testing for each gene based on the number of probes tested. The gray shaded box shows the 1.5-LOD support interval for the expansion propensity QTL based on tetranucleotide STRs. (B) Distribution of gene expression for strains with B versus D haplotypes. Panels show gene expression for each gene for strains assigned the B (red) versus D (blue) haplotypes at the QTL locus. Data shown are aggregated across all GeneNetwork data sets with a significant eQTL for each gene. Distributions per data set are shown in Supplemental Figure S20.

propensity than variants overlapping Msh3 or Atg10 (Fig. 1.2D; Supplemental Fig. 1.18). Overall, given its known role in STR stability and the high density of variants with predicted impact overlapping its mismatch recognition domain, our results provide compelling evidence for Msh3 as the gene driving this QTL.

1.4 Discussion

Genetic variation impacting proteins involved in DNA repair processes have the potential to drive genome-wide variation in mutation rates and patterns across individuals of a species, both in the context of disease but also across healthy individuals. Identifying these determinants may give insights into disease risk or progression and could improve population-genetic models of mutations. Recombinant inbred strains such as those in the BXD family have accumulated mutations over dozens of generations of inbreeding, offering a unique opportunity to map genetic determinants of these "mutator phenotypes." Here, we performed QTL mapping for three quantitative STR mutator phenotypes and identified a robust QTL on Chr 13 for expansion propensity in mice. The QTL region encompasses dozens of protein-coding genes, including *Msh3*, an important component of the DNA MMR machinery (Li 2008). We also identified two additional modest association peaks for the same phenotype (Supplemental Fig. 1.10). One of these overlaps a different MMR gene on Chr 17, *Msh5*, whose role in repeat expansions is less well studied. We did not identify signals at other genes well known to be involved in repeat stability, such as *Pms2* (Narayanan et al. 1997). This may be because of a lack of segregating functional variants in other relevant genes in this cohort or because of a lack of power to capture certain mutation events such as large expansions.

Definitively identifying a single causal gene or variant in the QTL locus identified is challenging in the BXD family, which harbors long unbroken haplotypes spanning several megabases (Ashbrook et al. 2021). The abundance of literature evidence regarding the role of *Msh3* in STR stability in other contexts, as well as the high density of variants in or near the key region of the protein important for recognizing mismatched DNA, strongly suggests it as a causal gene for this locus. However, we could not rule out a role for other genes in this region. In particular, *Atg10* falls closest to the center of the QTL peak, and eQTL signals for *Atg10* are most consistently colocalized with the QTL. We additionally identified multiple protein-coding variants and an SV overlapping this gene. However, *Atg10* has only been indirectly connected with DNA repair through the autophagy system (Gomes et al. 2017). Further, whereas *Msh3* is most highly expressed in spermatocytes and oocytes, where germline mutations are likely to arise, *Atg10* is most highly expressed in the heart and other structures less likely to be related to a mutator phenotype. We additionally identified high impact mutations in two genes not known to be involved in DNA repair (*Cmya5* and *Zcchc9*), but it is unclear how those would contribute to an STR mutator phenotype.

Msh3 is well known to be involved in regulating STR stability. *Msh3* is one of multiple homologs of the Escherichia coli MutS MMR protein, which recognizes mismatched bases in DNA that arise during DNA replication (Usdin et al. 2015). In mice and other eukaryotes, MutS proteins form two different heterodimers. MSH2 and MSH6 form MutSalpha, which primarily recognizes base substitutions and small insertion/deletion loops (IDLs) (Li 2008). MSH2 and MSH3 form MutSbeta, which recognizes long IDLs (Gupta et al. 2011), which often arise due to misalignment of strands at STR regions. Model organism studies have shown that both MutSbeta proteins MSH3 and MSH2 (Manley et al. 1999; López Castel et al. 2010), but not MutSalpha protein MSH6 (van den Broek et al. 2002), are required for the formation of pathogenic repeat expansions (Dragileva et al. 2009; Tomé et al. 2013a). This may result from

MSH3 stabilizing hairpin structures formed at repeats rather than repairing them (Mirkin 2007). On the other hand, germline defects in both MutSalpha proteins, but not MSH3 (Huang et al. 2001), are implicated in Lynch syndrome, a common cause of hereditary colon cancers characterized by high rates of MSI (Lynch et al. 2015). However, somatic mutations disrupting *MSH3* are often found in cancers showing MSI (Boland and Goel 2010). Specifically, *MSH3* deficiency has been linked to a type of MSI termed elevated microsatellite alterations at selected tetranucleotide repeats (EMAST) and to lower levels of MSI at dinucleotide repeats (Haugen et al. 2008; Campregher et al. 2012).

Naturally occurring sequence variants in *Msh3* have been shown to act as modifiers of the stability of CAG repeats in both mice and humans. Tomé et al. (2013a) identified multiple missense mutations in inbred mouse strains, including all four missense mutations between DBA/2J and C57BL/6J in exons 3 and 7 of *Msh3* identified in this study. They hypothesize that one of these, T321I, may destabilize the protein in DBA/2J. Consistent with our findings of increased *Msh3* expression and expansion propensity associated with B versions of *Msh3*, they showed that the C57BL/6J MSH3 protein variant is more highly expressed than the DBA/2J variant and is associated with increased CAG expansions compared with the MSH3 variant in BALB/cByJ mice, which share those same missense mutations with DBA/2J. Although we only considered RNA transcript levels here, which do not necessarily reflect protein levels, it was previously shown that *Msh3* transcript levels do reflect protein levels in mice (Tomé et al. 2013b). In humans, inherited variants in *MSH3* have been reported to modify the age of onset of Huntington’s disease (Wheeler and Dion 2021) and X-linked dystonia-parkinsonism (Laabs et al. 2021), presumably through modifying repeat stability, and *MSH3* is a current drug target of interest for Huntington’s disease (Kingwell 2021). Further, a polymorphism in the 5′ end of

MSH3 has been associated with increased *MSH3* expression and somatic instability of the trinucleotide repeat implicated in Huntington's disease (Flower et al. 2019).

Whereas previous studies of *Msh3* as a modifier of STR stability have focused on somatic variation at a small number of disease-associated loci, we report a novel association between sequence variants in *Msh3* and genome-wide germline mutation patterns at STRs. Our results suggest that in addition to these roles affecting somatic STR instability in disease, common mutations affecting *Msh3* may contribute to biases in mutation patterns in the germline at the hundreds of thousands of short STRs across the genome. The major signal identified was an association of the C57BL/6J version of *Msh3* with a higher propensity for STRs to expand. This association remained across a broad range of repeat lengths considered and was strongest for tetranucleotide STRs. On the other hand, we also found a modest increase in mutation rates in strains with the DBA/2J *Msh3* haplotype across all repeat unit lengths tested (2–4 bp), which was most prominent for longer repeats (parent allele length, >~30 bp). The expansion propensity and mutation rate results suggest a tradeoff in which too little *Msh3* may result in an MMR deficiency (as seen in EMAST) (Campregher et al. 2012), whereas increased *Msh3* activity results in more MMR activity but biases mutations toward expansions (as previously observed at the Huntington's disease and other repeats (Fig. 1.5; Wheeler and Dion 2021).

Similar to previous findings in inbred mice (Tomé et al. 2013a), we find evidence that both protein-coding sequence variants, as well as *Msh3* expression levels, could collectively contribute to the increased expansion propensity in mice harboring the B versus D haplotypes at this locus (Fig. 1.5). In addition to multiple protein-coding variants that have been previously reported (Tomé et al. 2013a), our analyses revealed a 387 bp indel near the 5′ end of the gene and falling between exons 4 and 5, which encode the DNA mismatch recognition domain. This

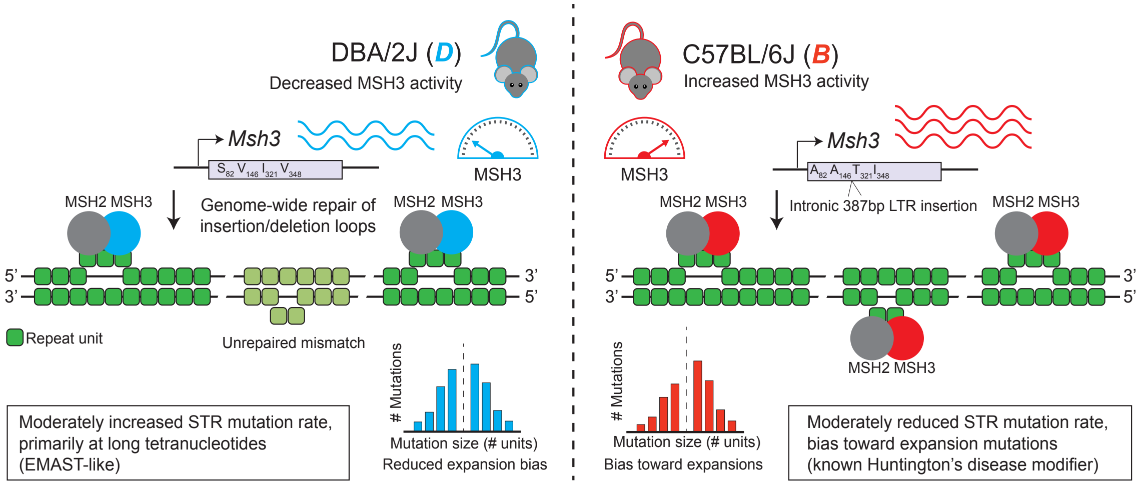


Figure 1.5 Schematic overview of proposed mechanisms for the expansion propensity QTL. BXD mice carrying the B haplotype (*right*) at the Chr 13 QTL locus tend to have higher *Msh3* expression than those carrying the D haplotype (*left*). The B and D *Msh3* variants also differ by four missense mutations (amino acid letter changes and positions are shown), as well as an intronic 387 bp LTR insertion only present on B (note the gene is not drawn to scale). MSH3 and MSH2 form the heterodimer MutSbeta, which recognizes strand misalignments, such as those formed by STRs (repeat units shown in green), across the genome during DNA replication. Mice with the D haplotype have slightly increased mutation rates, particularly at longer tetranucleotides, whereas mice with the B haplotype have reduced mutation rates but an increased propensity toward expansion mutations.

indel is owing to a partial intracisternal A particle (IAP) LTR insertion in C57BL/6J, which is missing in DBA/2J and many other classic mouse strains (Supplemental Fig. 1.28). IAP LTRs are one of the few active retrotransposon families in the mouse genome (Wang et al. 2019). Two of the most well-studied variants in mice have arisen through IAP LTR insertion: agouti viable yellow (Duhl et al. 1994) and Axin fusion (Vasicek et al. 1997). Although IAP LTR elements are typically heavily methylated (Walsh et al. 1998), the element at this locus is a member of the IAPLTR2a group. This group is overrepresented among hypomethylated LTRs (Ekram and Kim 2014), harbors transcription factor binding sites which can potentially contribute to regulation of nearby genes (Shimosuga et al. 2017), and has been shown to induce alternative splicing of nearby exons (Wang et al. 2019). Finally, this IAP element also forms an exon of a noncanonical transcript of *Msh3*, although it remains unclear if the NMD transcript is relevant to the expansion propensity phenotype. Although these sequence variants and the IAP could plausibly be causal drivers of the expansion propensity phenotype, we note experimental validation of individual causal genes or variants for this phenotype is challenging: The STR mutation phenotypes measured here are based on mutations that have arisen over decades of inbreeding and would not be evident in genome-edited cell lines or animals observed for a small number of generations.

Importantly, our study focused on germline mutations arising during parent-to-offspring transmission. Somatic mosaicism could not be assessed here, as we did not have available sequencing from different tissues of the same animal. Additionally, detecting somatic instability from a single WGS data set remains a difficult bioinformatics challenge and an important topic of future methods development. Notably, we do not directly assess parent-to-offspring mutation events as we focus on mutations that have already drifted to homozygosity in a particular strain.

Thus, the observed mutation sizes could have arisen as a result of numerous expansion and contraction mutations over time in some cases. This also means it is not possible to determine whether a particular mutation arose in the maternal or paternal germline. Although comparison of mutation patterns on sex chromosomes versus autosomes could give insight into a potential parent of origin effect, our analysis to assess this was underpowered owing to the low total number of observed sex chromosome mutations. It is known that germ lineages experience different processes of DNA metabolism compared with somatic tissues that differ between maternal and paternal lineages, and that these processes can alter STR mutation patterns (Pearson 2003). Our results are suggestive of a paternal effect, but future work is needed to more definitively assess this. *Msh3* is highly expressed in both male and female reproductive tissues, and we did not identify evidence of sex-specific expression patterns in other tissues. Thus, it is possible it could play a role in regulating STR mutations arising in both but is stronger in the male germline, in which frequent mitosis events present more opportunities for STR mutations to arise.

The fact that naturally occurring polymorphisms in the 5′ end of *Msh3* are associated with similar phenotypes in both humans and mice raises intriguing evolutionary implications and suggests polymorphism at this locus may confer a selective advantage. It is worth highlighting the interesting tradeoff noted above: Loss of *Msh3* may protect against expansions but, on the other hand, can result in MMR deficiencies, as seen in human cancers (Adam et al. 2016). On the other hand, increased *Msh3* expression can result in an increase of harmful expansions but could potentially protect against cancer. Interestingly, there is a significantly reduced prevalence of cancer among patients affected by Huntington’s disease and other repeat expansion disorders (Lucá et al. 2013; McNulty et al. 2018). Finally, it is possible that there is an advantage to keeping around a locus that promotes STR variability in general as a source of new and potentially adaptive changes upon which evolution can act (Kashi and King 2006). Although we did not assess the functional consequences of the new STR mutations, previous work has shown a role of STR variation in affecting gene expression and other phenotypes across multiple species (Vinces et al. 2009; Quilez et al. 2016; Fotsing et al. 2019). Leveraging the extensive phenotype information available for the BXD strains to perform detailed studies of the effects of STR variation on phenotype represents a rich area of future study.

In summary, our study reveals a novel QTL for STR mutation patterns, providing a striking example of the influence of inherited variants on germline mutation properties. Beyond *Msh3*, additional modifiers for both STR and other mutator phenotypes are likely to exist in humans or in other model organism data sets. We anticipate that further investigation of these mutation modifiers will provide new insights into mutation processes both in health and disease.

1.5 Methods

1.5.1 WGS and variant calling in the BXD cohort

Genome-wide STR and SNP genotypes for males from 152 RI strains and the two BXD founders, C57BL/6J (B) and DBA/2J (D), were previously generated from WGS data based on the 10x Chromium system (see "Data access"). The origin tissues for the samples were spleen and tail. For clarity, the STR genotyping process is summarized below.

We used Tandem Repeats Finder (Benson 1999) to identify regions within the mm10 mouse reference genome predicted to harbor STRs with repeat unit lengths up to 20 bp. We used GangSTR (Mousavi et al. 2019) to genotype the reference STR loci in 152 BXD strains and the two founder strains, C57BL/6J and DBA/2J. The 10x Chromium workflow requires a large amount of PCR amplification, which can introduce significant "stutter" errors in repeat copy number at STR regions, particularly at dinucleotide repeats (Ashbrook et al. 2022). To reduce the effects of these stutter errors, we first used HipSTR (Willems et al. 2017) to perform per-locus stutter estimation. We then called GangSTR separately on each strain using our STR reference panel, trimmed and dedupped reads, and per-locus stutter error probabilities as input. A custom build of GangSTR was used to handle unequal read lengths present in the BXD Chromium data (https:// github.com/gymreklab/GangSTR/tree/fix_read_length). STR genotypes for each strain were filtered using dumpSTR (Mousavi et al. 2021) v1.0.0 with the options --min-call-DP 20 --max-callDP 1000 --min-call-Q 0.9 --filter-badCI --require-support 2 --readlen 128 to remove genotype calls with insufficient read depth, read support, or quality scores. Calls were then merged into a single multisample VCF file containing maximum likelihood diploid genotypes for each STR in each strain. The merged VCF was further filtered to remove (1) STRs overlapping known segmental duplication regions in the mm10 reference based on the mm10.genomicSuperDups table obtained from the UCSC Table Browser (Karolchik et al. 2004), (2) STRs with calls in less than 50 unfiltered strains, (3) STRs with no variation in repeat number across all strains, and (4) STRs for which variants from the mm10 reference were only observed in heterozygous genotypes. Full details of the genotyping pipeline are described by Ashbrook et al. (2022). STR genotyping was performed here for Chr X and Chr Y using an identical pipeline as for autosomes, with the exception that we required a minimum DP of 10 (rather than 20) due to the lower coverage on the sex chromosomes.

Epoch labels and number of generations of inbreeding were obtained from Supplemental Table S1 of Ashbrook et al. (2022). For epoch 7 strains (BXD221–BXD227), which followed a more complex breeding structure, we used the number of inbreeding generations after mating two previously inbred parental BXD strains.

1.5.2 SNP marker maps for founder inference and interval mapping

We prepared a marker-by-strain matrix of founder labels (B vs. D) for BXD strains using SNP genotypes at 7,124 autosomal LD-pruned markers published on GeneNetwork (http://gn1.genenetwork.org/webQTL/main.py?FormID=sharinginfo&GN_AccessionId=600). For SNPs not directly genotyped from WGS in the BXD, we chose the next closest SNP based on genomic distance that was <500 kbp away. In a small number of cases, the closest SNP was the same for multiple markers, in which case a single marker/SNP combination was retained producing a final list of 7,101 markers. R/QTL2 (Broman et al. 2019) version 0.24 was used to calculate founder genotype probabilities suitable for QTL mapping using the "calc_genoprob" function with default parameters. We then generated a complete list of SNP founder labels with maximum marginal probabilities using the "maxmarg" function with "minprob" parameter set to 0.5. Founder labels at individual markers were used to find start and stop positions of haplotype blocks using a connected components clustering approach (R tidygraph) (R Core Team 2021; https://cran.r-project.org/web/packages/tidygraph/index.html).

1.5.3 Identifying and phasing new STR mutations

We identified candidate STR mutations as STR genotypes in BXD strains not matching genotypes in either of the two founder strains. In cases in which one or both founders were not directly genotyped, we first inferred missing STR calls in founders (below). We intersected each candidate new mutation with haplotype blocks inferred from SNPs to assign each mutation as occurring on the *B* versus *D* haplotype. STRs falling in a gap between blocks were assigned to the nearest block. We excluded new variants in which either the BXD or founder strain was heterozygous, which likely indicates either poor quality STR genotypes or incomplete inbreeding at that locus. Finally, we excluded strain BXD194, in which we found an outlier number of new mutations (more than twofold higher than other strains in the same epoch) from downstream analyses.

1.5.4 Inferring missing founder STR genotypes

We used R/QTL2 to infer missing founder STR genotypes from genotypes observed in BXD strains. First, we imputed founder labels (*B* or *D*) for each STR genotype in the BXD strains. For the subset of loci at which both founder strains were genotyped and did not share a common allele, we could unambiguously assign *B* or *D* genotype labels to each genotyped BXD strain. BXD strains with genotypes not matching either founder were assigned missing labels. For the remaining polymorphic loci missing at least one founder genotype, we could not directly infer the founder label and initially set all genotypes at those loci to missing values. We used the R/QTL2 "interp_map" function to interpolate linkage distances between STRs from physical and genetic SNP marker maps at the 7,101 LD-independent markers described above. We then used R/QTL2 functions "calc_genoprob" followed by "maxmarg" to impute missing founder labels. Then, for each STR with a missing founder genotype, we determined the distribution of repeat lengths in strains inferred to have the corresponding founder label at that locus. If at most one de novo genotype was present at the locus and if the majority of BXD strains had the founder genotype, the founder was inferred to have the modal allele. Otherwise, the locus was removed from downstream analysis.

1.5.5 Characterization of new STR mutations

We performed PCA to characterize sharing of new mutations across strains. We constructed a strain-by-locus matrix of indicator values indicating the presence (one) or absence (zero) of a new STR genotype in each strain at each locus. We then performed PCA using the builtin "prcomp" function in R with centering but without scaling.

1.5.6 Computing STR mutator phenotypes

We calculated three separate mutator phenotypes for each strain. *Mutation count* was calculated as the number of STRs with new mutations divided by the number of successfully genotyped loci in that strain. *Mutation size* was calculated as the average difference in repeat count between the new genotype and the founder genotype at each mutation. Mutation size was computed separately for expansion and contraction mutations. *Expansion propensity* was calculated as the fraction of new mutations in each strain for which the RI genotype was longer than the founder genotype. Unless otherwise noted, we removed STR mutations seen in more than 10 strains, as those likely do not represent new mutations.

1.5.7 QTL mapping for STR mutator phenotypes

QTL mapping for each mutator phenotype was performed based on the set of LD-pruned SNPs described above using a linear mixed model approach implemented in R/QTL2. Each phenotype was analyzed separately. We used the "calc_kinship" function to prepare a strain relatedness matrix using the leave-one-chromosome-out (LOCO) method. In addition to supplying a vector of phenotype values, genotype probabilities, and kinship matrices, we also input a vector of the number of inbreeding generations as a covariate. We used "scan1perm" to calculate permutation-based genome-wide significance thresholds based on 100 permutations. For each QTL analysis performed, strains with fewer than 10 total new mutations were excluded from analysis because they produce noisy mutator phenotype values.

1.5.8 Variant annotation

The initial set of variants for annotation analysis contained 66,017 SNPs and 1,040 STRs genotyped previously in the BXD cohort (Ashbrook et al. 2022) and located between the boundaries of the confidence interval for the QTL on Chr 13. We additionally obtained genotypes for 8,649 SVs based on pangenome analysis (see below). After filtering for variants within protein-coding genes in the QTL region based on the GENCODE M25 release gene annotations, 35,031 SNPs, 576 STRs, and 4,135 SVs remained. SVs <50 bp were removed, leaving 983 SVs. After filtering for only segregating variants and removing variants in which more than half the strains had a missing value, 5,982 SNPs, 214 STRs, and 733 SVs remained. The non-major-allele frequency was calculated for each variant as the proportion of alleles at the locus that were not the most abundant allele after removing strains with missing genotypes. We used VEP (McLaren et al. 2016) v103.1 with the Ensembl cache v102 to predict the impact of each variant. VEP assigns one of four IMPACT ranks (high, moderate, low, and modifier) along with predicted consequences to each variant overlapping a transcript or a regulatory feature. The strength of association between the genotype at each variant and the expansion propensity phenotype was taken as the one-sided P-value of the F-statistic for an ANOVA model with genotype as a categorical predictor variable using R. Twenty-four SV loci were filtered out because of not returning an association value, for a final count of 9,103 SNPs, 160 STRs, and 959 SVs. There was an average of 4.3 transcripts and 10.5 regulatory features per gene, for a total of 328 features and 25,746 variant-feature pairs. The variant-feature pair with the most severe impact and consequence was selected among variants predicted to have multiple consequences and/or impacts on protein features.

1.5.9 Pangenome analysis of SVs

The BXD pangenome for Chr 13 was built from data of 148 strains (four strains were excluded because of poor assembly quality) using haploid assemblies of 10x reads obtained by Supernova (Weisenfeld et al. 2017). To restrict the analysis to Chr 13, haploid assemblies were mapped against the GRCm38/mm10.fa reference genome using wfmash v.0.6.0

(https://github.com/waveygang/ wfmash; https://doi.org/10.5281/zenodo.6949373). Only assemblies mapping to Chr 13 were used to build the pangenome with pggb (Garrison et al. 2023) v0.2.0 using the following combination of parameters: pggb-0.2.0 -i chr13.pan+ref.fa.gz -o chr13.pan+ref -t 48 -p 98 -s 100,000 -n 140 -k 229 -O 0.03 -T 20 -U -v -L -Z.

Regions of the pangenome with depth < 10× were removed using odgi (Guarracino et al. 2022). Variant calling from the pangenome was performed with vg (v1.35.0-59-ge5be425c6) (Garrison et al. 2018) using the following combination of parameters: vg-e5be425 deconstruct -t 16 -P REF -e -a -H "#" graph.gfa > graph.vcf.

The variant call set was processed to remove missing data, sites where alleles are stretches of Ns, homozygous reference genotypes, and variants <50 bp and >10 kbp before normalization and decomposition using BCFtools (Bonfield et al. 2021) under standard parameters. The resulting VCF file was visualized using bandage v0.8.1 (Wick et al. 2015).

Reference and alternate allele sequences for SVs were extracted from the resulting variant call file using "bcftools query." Each alternate sequence was then aligned to the reference using the Needleman–Wunsch global pairwise alignment implemented in the "pairwiseAlignment" function from the Biostrings v2.60.1 R package. This allowed for splitting complex SV sequences spanning multiple kilobases into smaller individual insertions/deletions for variant effect analysis. We removed singleton variants and those <50 bp in length.

1.5.10 eQTL analysis

We generated a list of 264 expression data set files available from GeneNetwork's interplanetary file system (IPFS) using the "lftp" tool. Of these, 242 data sets contained BXD strain data. Some GeneNetwork data sets do not reflect the nomenclature change of the

BXD24/BXD24_Cep sister strains. To avoid ambiguity and standardize strain names with newer data sets, BXD24 and BXD24a were relabeled as BXD24_Cep and BXD24, respectively, in the data sets GN267, GN373, GN385, GN410, and GN414, which contained expression values for both of these strains. Similarly, BXD24a was relabeled as BXD24 in the data sets GN274, GN275, GN302, GN308, GN325, GN374, GN375, GN387, and GN702. Probe information and per-strain gene expression values were extracted into separate tables of a sqlite3 database to facilitate querying. Probes with missing genomic location information were removed. Finally, probe coordinates were converted from the mm9 to the mm10 reference using the UCSC Genome Browser liftOver tool (Hinrichs et al. 2006), and probes that failed remapping to the new reference were discarded.

Each GeneNetwork data set represents a distinct processing configuration of data generated from an experimental study. Processing steps include signal intensity normalization, strain and probe filtering, rescaling, and correction of batch effects. Multiple data sets may be available for studies in which both gene- and exon-level data have been collected. Further, study data may be split up into multiple data sets according to the sex of the animals or by treatment group such as diet or drug exposure. To avoid overcounting, we selected a single representative data set using a heuristic approach to make the selection based on strength of signal and processing conditions. Exon-level data were preferred to gene-level data due to increased probe density. More recently reprocessed data sets were preferred to older ones. Data from control groups were preferred to data from experimentally treated groups. Combined male and female data were preferred to sex-specific data. Data sets with more strains were preferred to data sets with fewer strains. A summary of selected and available data sets for each study is available in Supplemental Table S5.

We then queried expression values for all probes falling within the expansion propensity QTL region on Chr 13 in each data set. GN227 lacked probe data in this region and was excluded. Probe mapping information was either taken directly from the GeneNetwork data set or queried from Ensembl's BioMart data mining tool release 102 using the biomaRt (Durinck et al. 2009) R package. Unmapped probes were removed from analysis. We then checked whether probe coordinates were contained within the start and stop positions of each probe's corresponding gene and removed those that did not. For each Affymetrix probeset representing a collection of probes, we used the UCSC (Kent et al. 2002) BLAT tool to find the matching genomic location of individual probe sequences. We discarded probe sets in which any contained probe did not match within the coordinates of its assigned gene. We then used probe coordinates to calculate the number of segregating variants that each probe overlapped using the "bedtools intersect" command available from the BEDTools (Quinlan 2014) package. Additionally for each probe, we calculated the number of variants at which each strain differed from the mm10 reference, which represents the number of mismatches an array probe would be expected to have when hybridizing with a DNA library sample from a given strain. We then performed eQTL mapping on Chr 13 using the same set of LD-independent loci and kinship matrix. The covariate vector from the QTL mapping was supplemented with the number of expected hybridization mismatches for each probe/strain combination to account for the expected differences in hybridization efficiency. The number of strains per data set ranged from 11 to 79. For comparison, we remapped the mutation propensity phenotype using only strains available in each of the gene expression data sets. Monoallelic markers conditioned on the subset of strains available in each expression data set were removed.

Notably, it is common for multiple microarray probes (probe sets) to target the same gene, especially for exon-based microarrays. We observed high variability for gene expression measurements between probes targeting the same gene in a given data set. To limit the rate of false eQTL signal discovery, we applied the Benjamini–Hochberg multiple hypothesis testing correction (Benjamini and Hochberg 1995) to the vector of peak LOD values for each gene–data set pair. We selected a representative probe for each gene having the highest adjusted peak LOD value within the expansion phenotype QTL region on Chr 13 for gene-level analysis. For visualization of eQTL traces, LOD values at each marker were scaled by the ratio of the peak adjusted LOD to the peak LOD for each gene.

1.5.11 Genomic data for classic mouse strains

Read alignment BAM files for the common laboratory mouse strains—129S1/SvImJ, NZO/HlLtJ, NOD/ShiLtJ, CAST/EiJ, PWK/ PhJ, A/J, and WSB/EiJ—were downloaded from the Mouse Genomes Project ftp server hosted at ftp://ftp-mouse.sanger.ac .uk/current_bams. Variant call files for these strains were similarly queried from ftp://ftp-mouse.sanger.ac.uk/current_snps.

1.5.12 Tissue-specific expression of DNA repair genes

Tissue-specific expression of *Msh3* and other DNA repair genes (Supplemental Table S6) was obtained from the Bgee database (Bastian et al. 2021), accessed on November 7, 2022.

1.6 Data access

WGS data and genotype calls for 152 strains from BXD were generated previously (Ashbrook et al. 2022) and are available on the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ ena/browser/home) under accession number PRJEB45429). STR genotypes are available on the European Variation Archive (EVA; https://www.ebi.ac.uk/eva/)

under accession number PRJEB61080. The set of new mutations and STR loci included in this analysis are available in Supplemental Datasets S1–S3. Workflow and analysis scripts are available at GitHub (https:// github.com/gymreklab/BXD-STR-Mutator-Manuscript) and as Supplemental Code.

1.7 Supplementary Figures

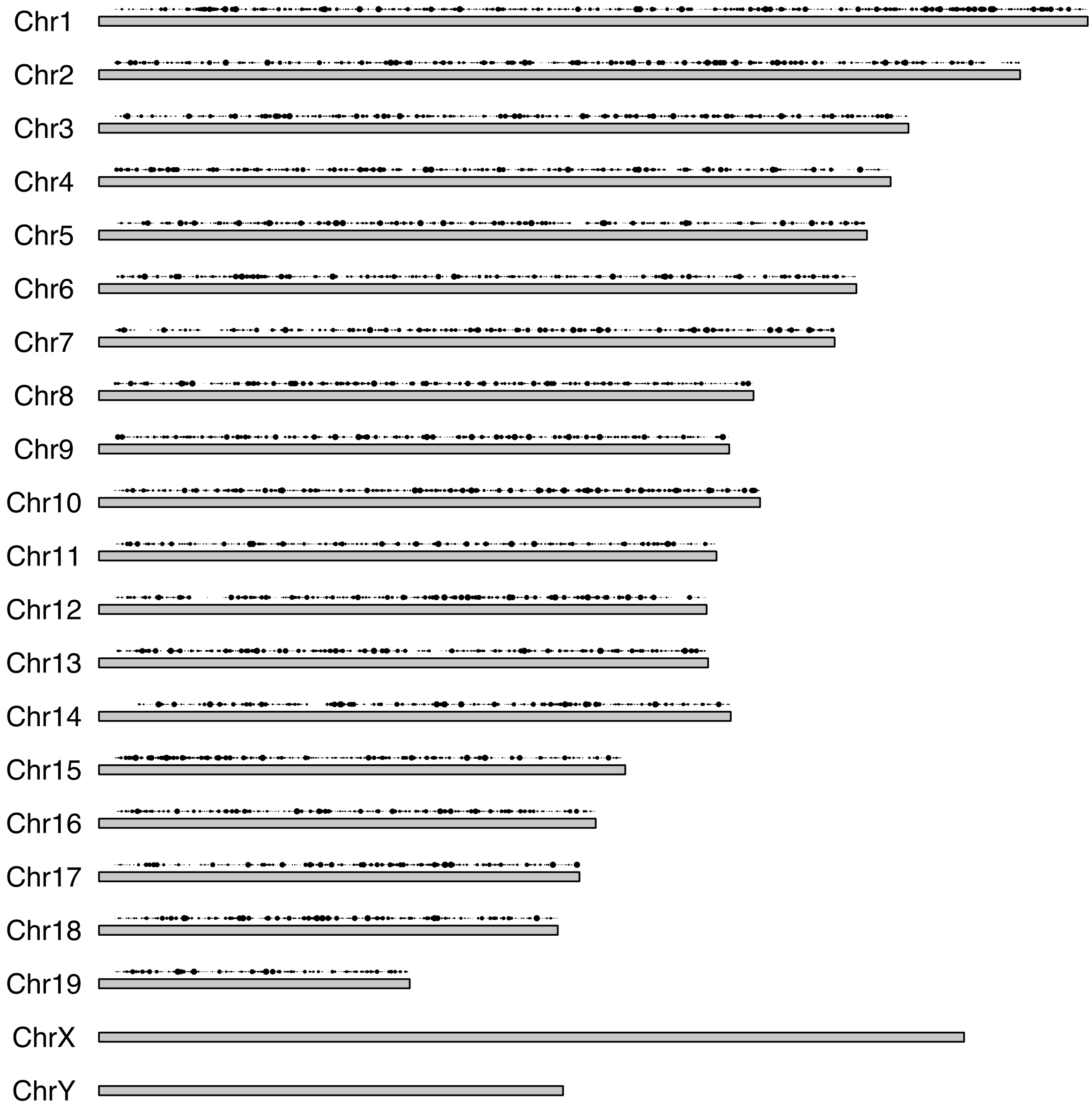


Figure 1.6 Localization of new autosomal mutations at STRs. Each dot represents a single STR for which at least one new mutation was observed. The size of each dot scales with the number of strains for which a mutation was observed at that locus. Loci at which more than 10 new mutations were identified were filtered. Dot sizes range from 1-10 mutations. Plots were made with the karyoploteR (Gel and Serra 2017) package.

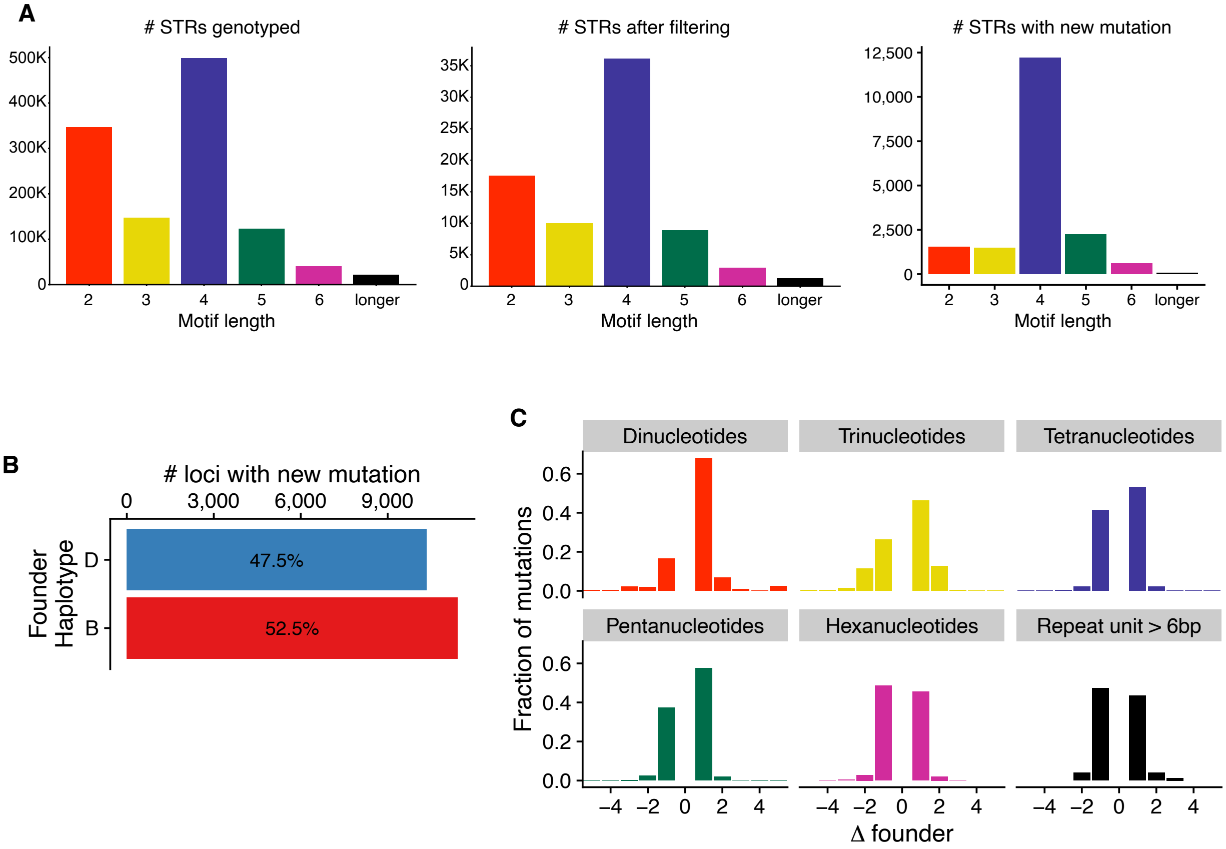


Figure 1.7 Summary of new STR mutations in BXD. A. Distribution of repeat unit lengths. The number of new mutations at STRs with each repeat unit length (bp) is shown (left=all genotyped STRs, middle=all STR loci passing initial filtering, right=all STRs with new mutations). **B.** Distribution of the founder haplotypes for new mutations. Bars show the number of new STR mutations occurring on “B” (red) vs. “D” (blue) founder haplotypes. **C.** Distribution of mutation sizes for each repeat unit length. The x-axis shows mutation sizes in terms of the difference in number of repeat units (RU) from the founder allele. Positive sizes indicate expansions and negative sizes indicate contractions.

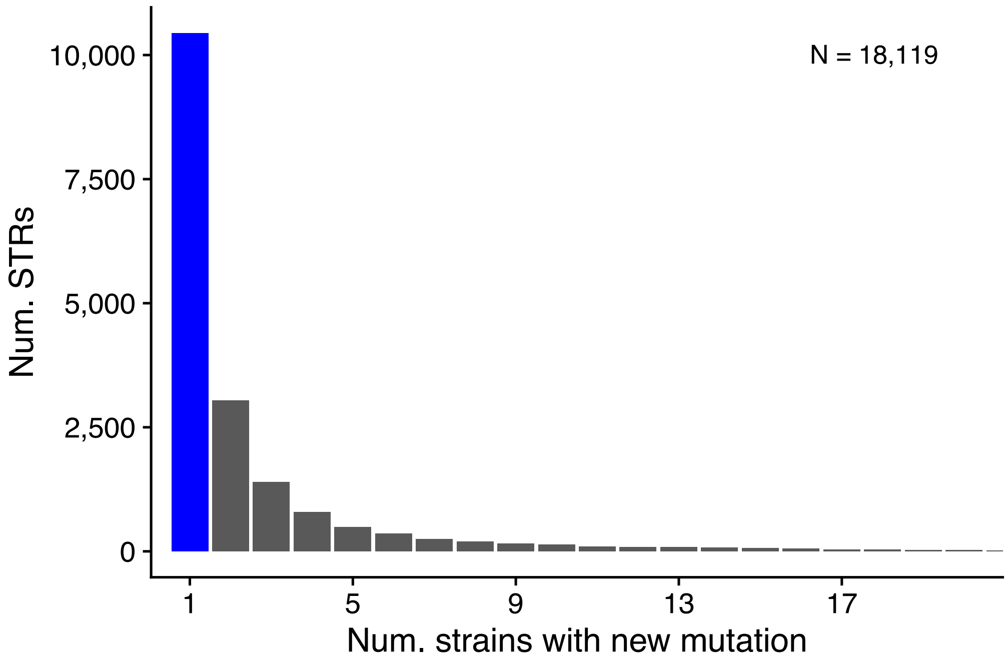


Figure 1.8 Distribution of the number of strains carrying the new allele at each of the STRs for which at least one new mutation was identified. Singleton mutations, seen only in a single strain, are shown in blue.

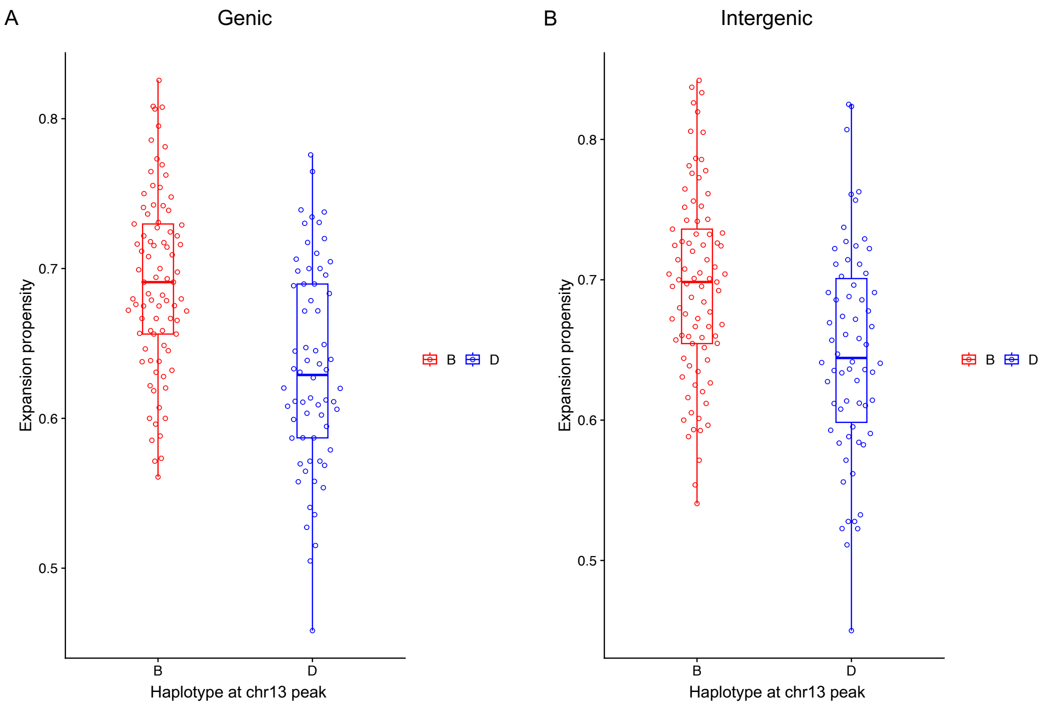


Figure 1.9 Expansion propensity phenotype at the Chr 13 QTL for mutations in genic vs. intergenic regions. Each point represents one strain. We used SNP haplotype blocks to assign each strain as harboring either the B (red) or D (blue) haplotype at the Chr 13 locus. The y-axis denotes expansion propensity computed based on STR mutations occurring in either genic **(A)** or intergenic **(B)** regions.

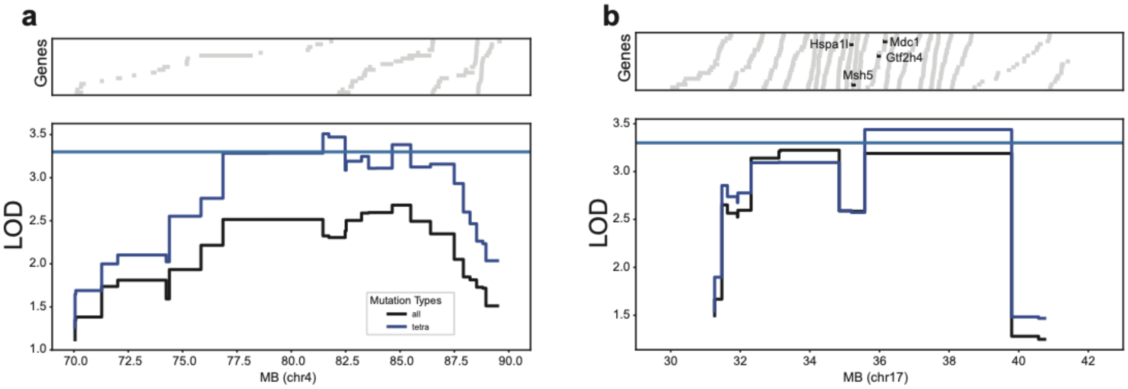


Figure 1.10 Genes located in or near the modest QTL peaks for expansion propensity. The y-axis shows the QTL signal (LOD score) for expansion propensity. Black line=all STRs, blue line=tetranucleotide STRs only. Horizontal bars denote genes near the center of the QTL peak. Genes known to be involved in DNA repair are highlighted. The peak on Chr 4 does not overlap any known DNA repair genes. There are 76 genes shown for the Chr 4 region and 371 genes shown for the Chr 17 region.

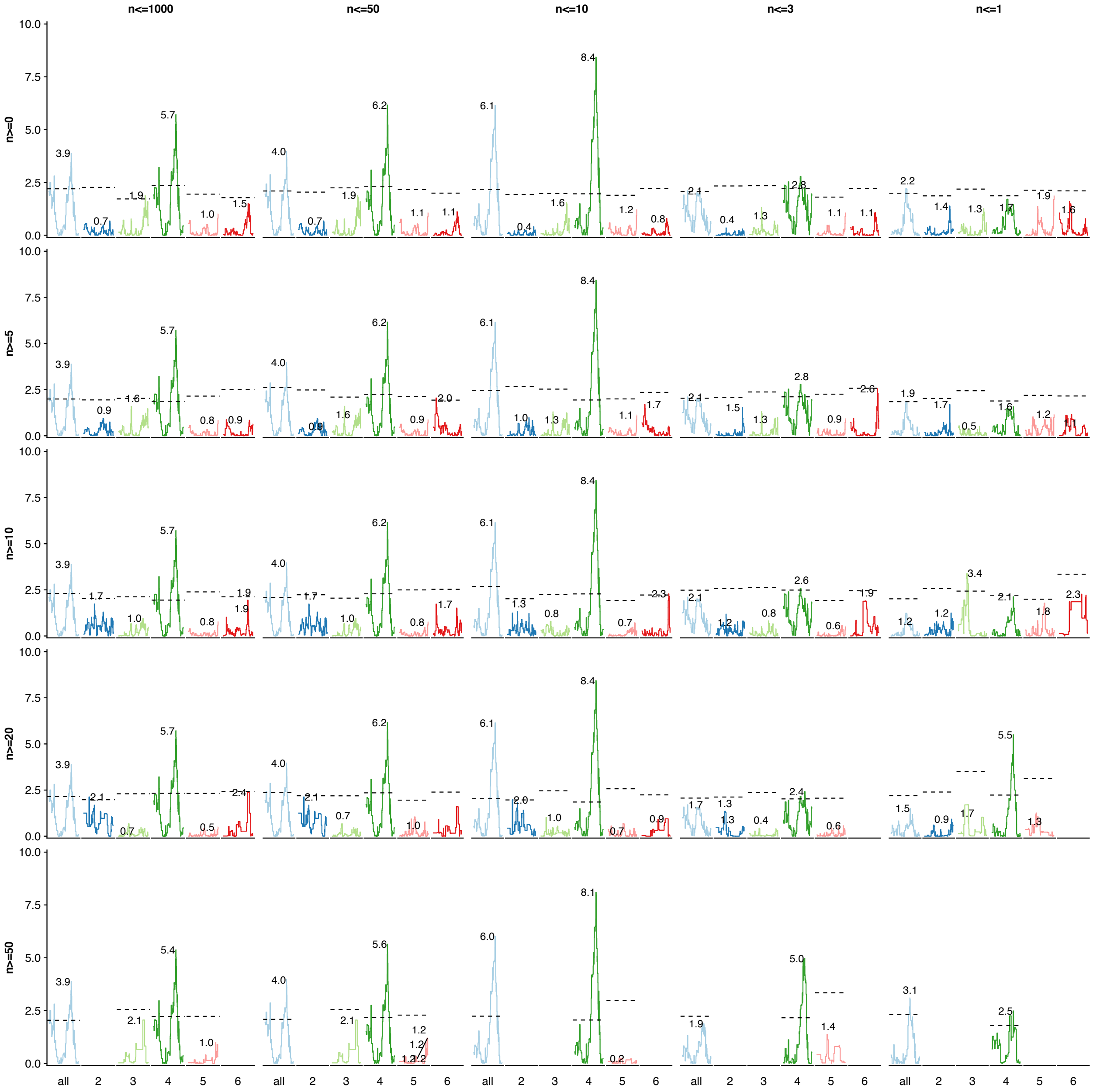


Figure 1.11 Evaluating robustness of the Chr 13 association signal for expansion propensity. In each panel, the x-axis denotes the repeat class (from left to right: all STRs, and including only STRs with a repeat unit length of 2-6bp). Within each class in each panel, the x-axis denotes genomic location on Chr 13 and the y-axis denotes logarithm of the odds (LOD). The max LOD is annotated for each class. Each row denotes a different threshold for the minimum number of new STR mutations for a strain to be included in the analysis (strain filtering). Each column denotes a different threshold for filtering the maximum number of strains a particular new STR mutation could be observed in (frequency filtering). Dashed horizontal lines represent permutation thresholds for genome-wide significance in each class. Overall, strain filtering has little effect whereas frequency filtering indicates the association signal is restricted to relatively new mutations. In all cases, tetranucleotides, the largest STR class in our dataset, show the strongest signal.

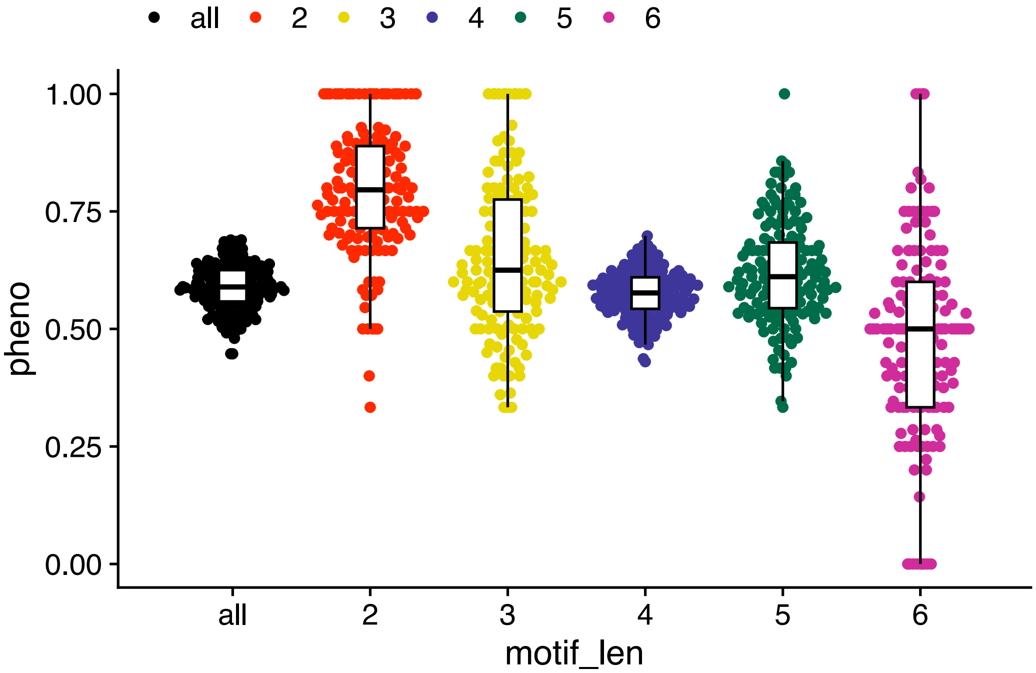


Figure 1.12 Distribution of expansion propensity for each strain for different repeat classes. Expansion propensity was computed separately considering only STRs with repeat units of a specified length (black=all STRs; red=dinucleotides; gold=trinucleotides; blue=tetranucleotides; green=pentanucleotides; purple=hexanucleotides).

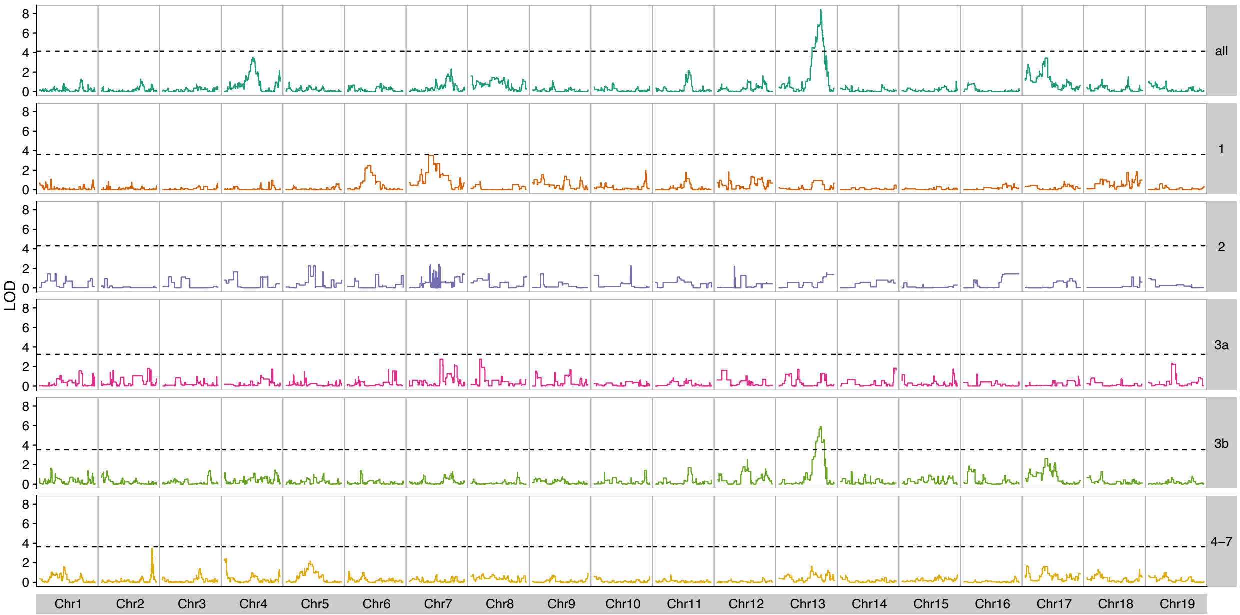


Figure 1.13 Expansion propensity QTL mapping in each BXD epoch. We repeated QTL mapping separately using only strains in each epoch and including only tetranucleotide loci. Each row represents a different epoch. In each row, the x-axis denotes genomic location and the y-axis denotes LOD score. Permutation based thresholds are shown as dashed horizontal lines.

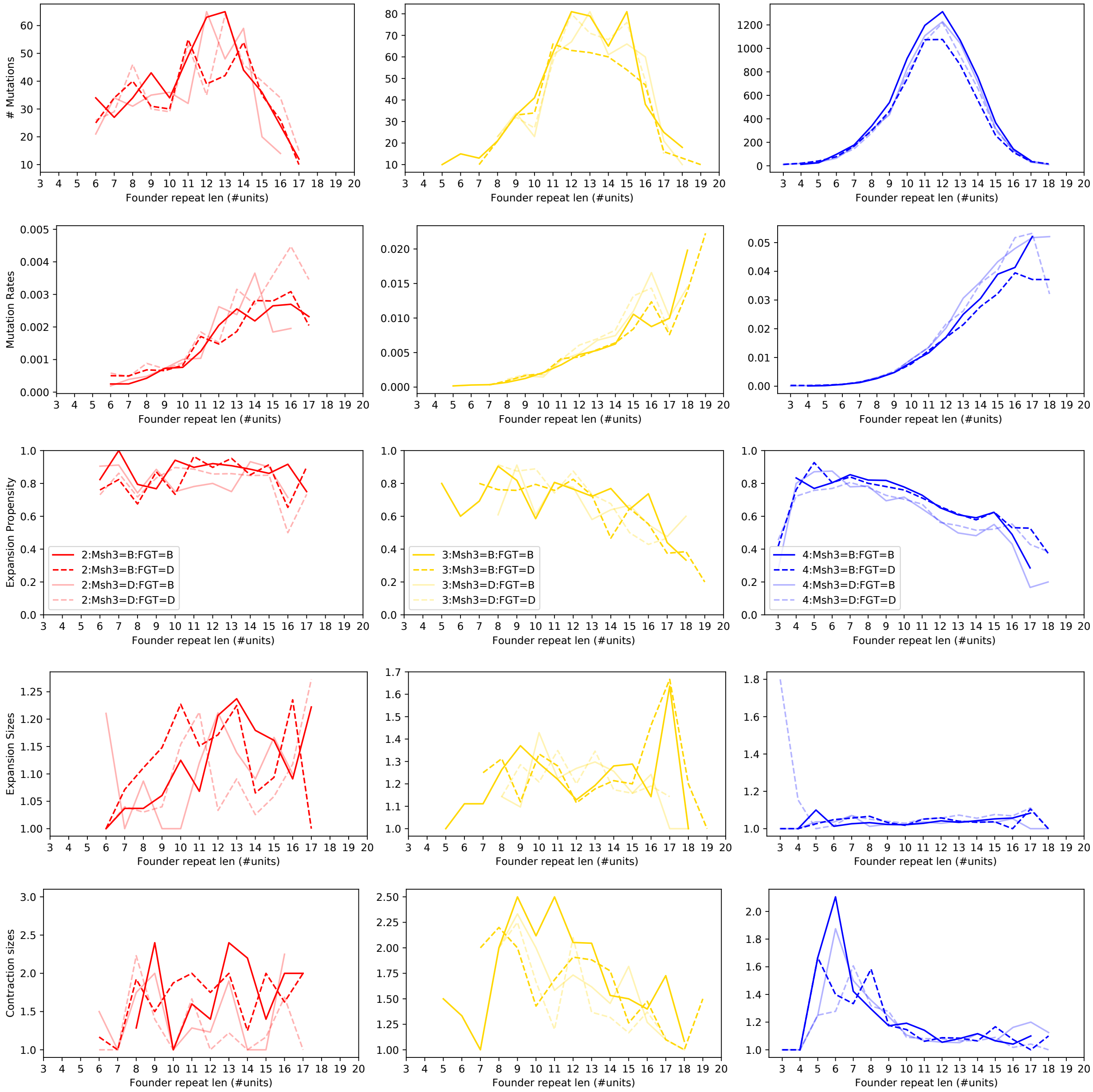


Figure 1.14 Overview of STR mutation patterns. In each panel, the x-axis gives the founder repeat length, based on the inferred founder haplotype at each STR locus (see **Methods**). Each row shows a different mutation metric, and each column is for a different repeat unit length (left=dinucleotides, middle=trinucleotides, right=tetranucleotides). In each panel, dark lines indicate patterns in strains which inherited the B haplotype at the Chr 13 QTL locus and shaded lines indicate patterns in strains with the D haplotype. Solid lines show data for STRs inherited on a local B haplotype and dashed lines are for STRs inherited on a D haplotype (e.g. as in the toy example in **Fig. 1.1A**). Rows, starting from the top, show the following metrics: (1) Total number of mutations observed in each category, (2) Relative mutation rate, computed as the number of mutations divided by the number of non-missing genotype calls in each category, (3) Expansion propensity of mutations in each category, (4) Mean size of expansion mutations, and (5) Mean size of contraction mutations.

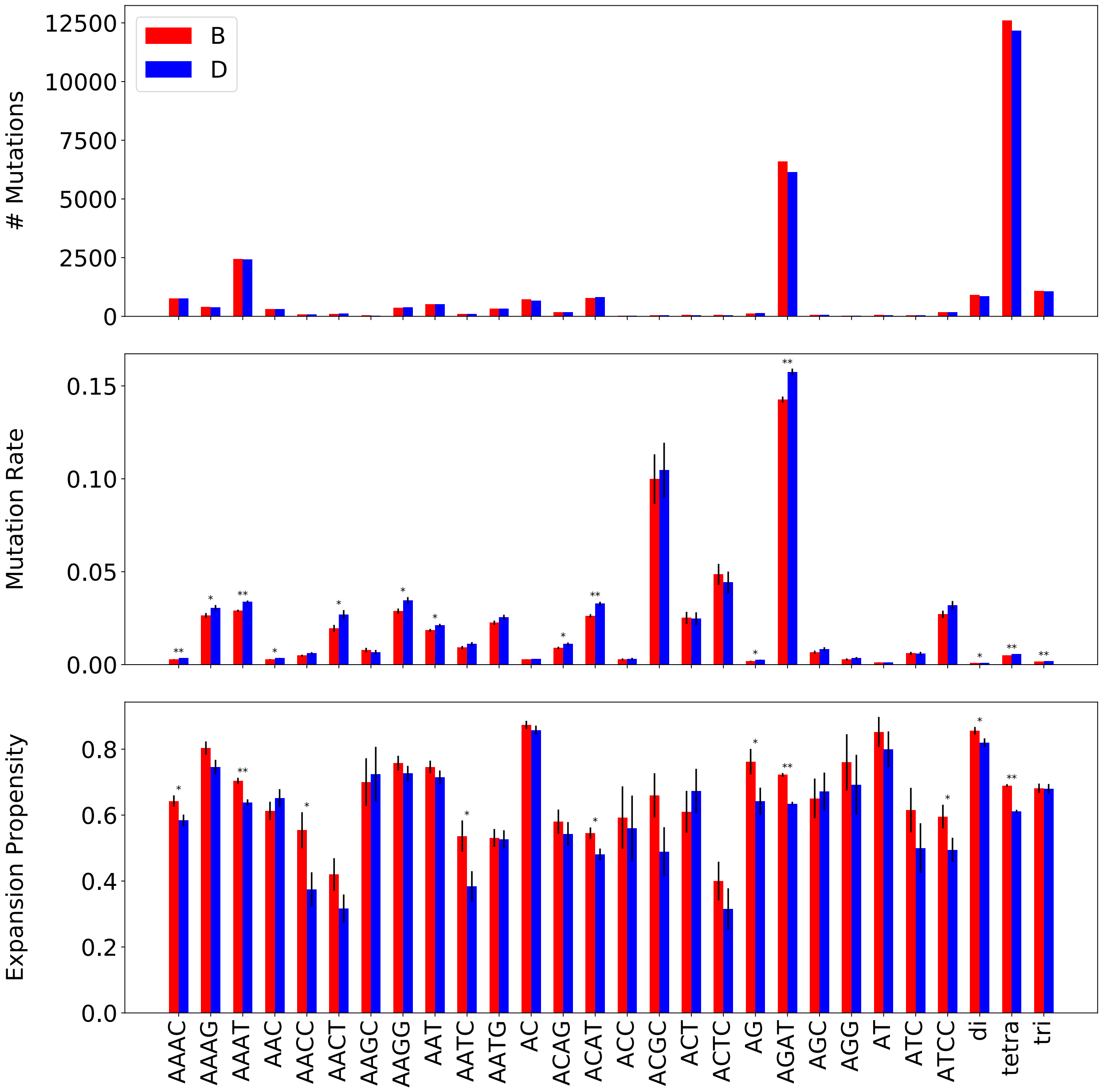


Figure 1.15 Mutation patterns stratified by repeat unit and haplotype at the Chr 13 QTL peak. In each panel, the x-axis denotes repeat units. Dark red=B haplotype and dark blue=D haplotype at the Chr 13 peak. ** denotes p<0.05 after Bonferroni correction and * denotes nominal two-sided z-proportions test nominal p<0.05. **Top:** The y-axis gives the total number of mutations observed for each repeat unit. **Middle:** The y-axis denotes relative mutation rate computed as the average number of mutations per strain divided by the total number of genotyped loci in each category. **Bottom:** The y-axis gives the percent of mutations for each repeat unit that are expansions.

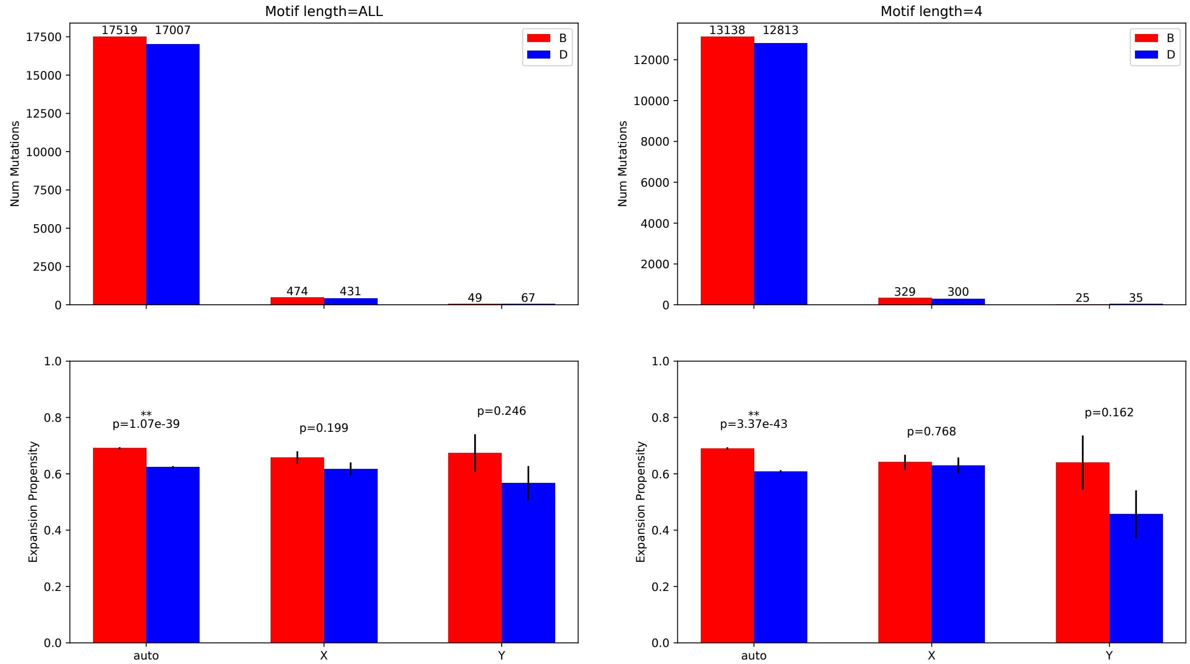


Figure 1.16 Mutation patterns by haplotype at the Chr 13 QTL peak for autosomes, Chr X, and Chr Y. In each panel, the x-axis denotes the chromosome type STR mutations occur on: autosomes, X, or Y. Dark red=B haplotype and dark blue=D haplotype at the Chr 13 peak. **Top:** The y-axis gives the total number of mutations observed for each category, after filtering mutations occurring in more than 10 strains. **Bottom:** The y-axis gives the percent of mutations for each repeat unit that are expansions. Bottom plots are annotated with the p-value from a two-sided z-proportions test. ** denotes p<0.05 after Bonferroni correction and * denotes nominal p<0.05. Left plots are computed based on all STRs, and right plots are computed based only on tetranucleotide STRs.

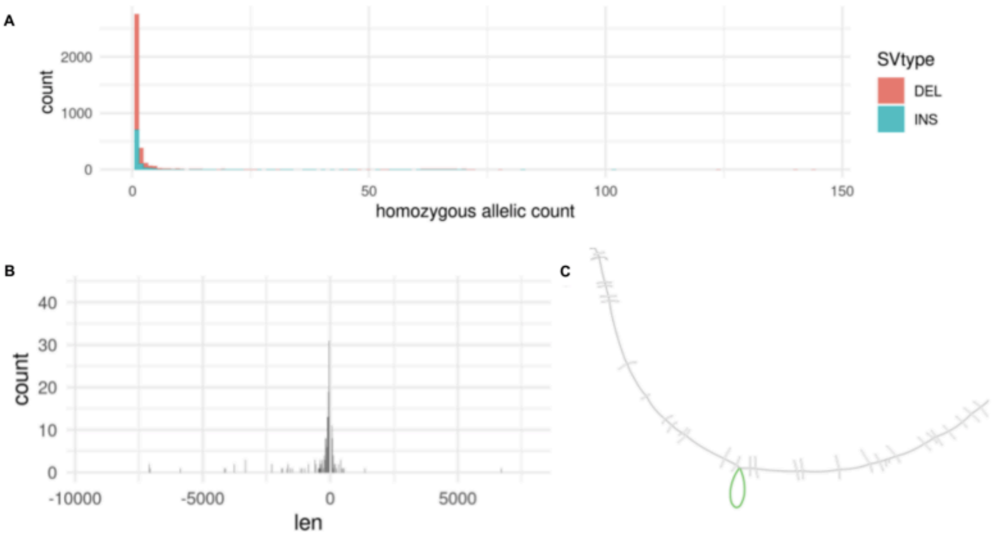


Figure 1.17 Features of the SVs discovered from pangenome analysis of Chr 13. A. Allele frequency spectrum of the 3,698 SVs with length >50bp and <10kbp in a region encompassing the QTL of interest on Chr 13. **B.** Distribution of the length of insertions and deletions. **C.** Bandage (Wick et al.) representation of the candidate region on Chromosome 13 (mm10, chr13:92,345,000-92,351,498) containing the 387bp insertion found in the 66 mice with the C57BL/6J background for that region.

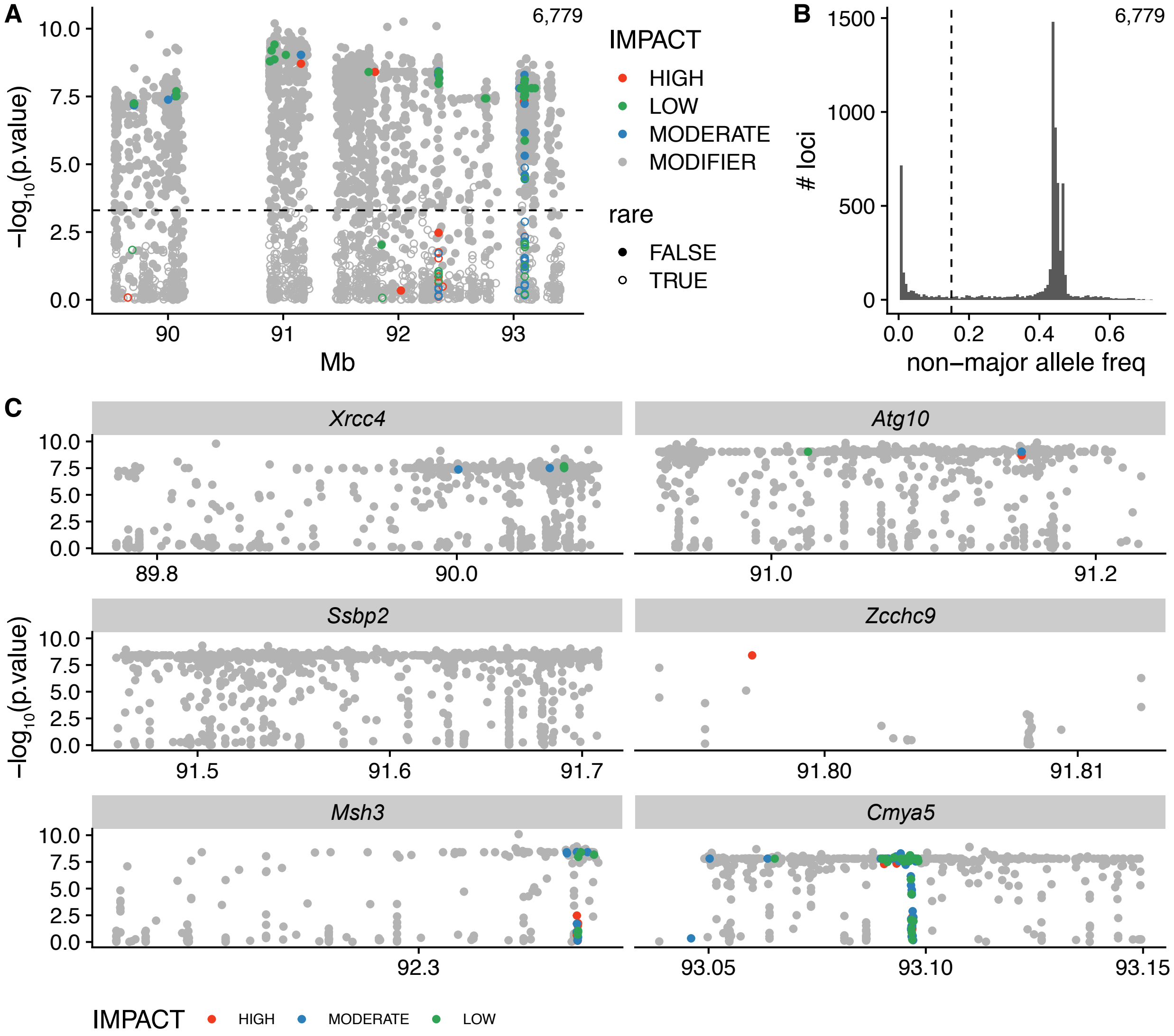


Figure 1.18 Annotation and selection of impactful variants within genes in the Chr 13 QTL for expansion propensity. For plots in A and C, the x-axis gives the genomic coordinate and the y-axis gives the association (-log10 p-value). Each dot represents a variant, and variants are colored by their impact predicted by VEP (red=high, blue=moderate, green=low; gray=modifier). **A. VEP-annotated variants across the entire QTL region.** Most annotated variants are located in intronic regions and have a predicted "modifier" impact. Weakly associated variants were removed from further analysis using a threshold of 3.3 on the association statistic (dashed horizontal line) as suggested on the GeneNetwork website (http://gn1.genenetwork.org/glossary.html). Filled dots represent common and empty circles represent rare variants based on the threshold identified in panel B. **B. Distribution of non-major allele frequencies.** Rare variants with an artificially strong association statistic due to overleveraging of outliers were removed using a threshold (dashed vertical line) on non-major allele frequency. **C. Detailed view of VEP-annotated variants.** Views are shown for genes known to be involved in DNA repair (Xrcc4, Atg10, Ssbp2, *Msh3*) or genes for which high-impact variants were detected (Cmya5, Zcchc9).

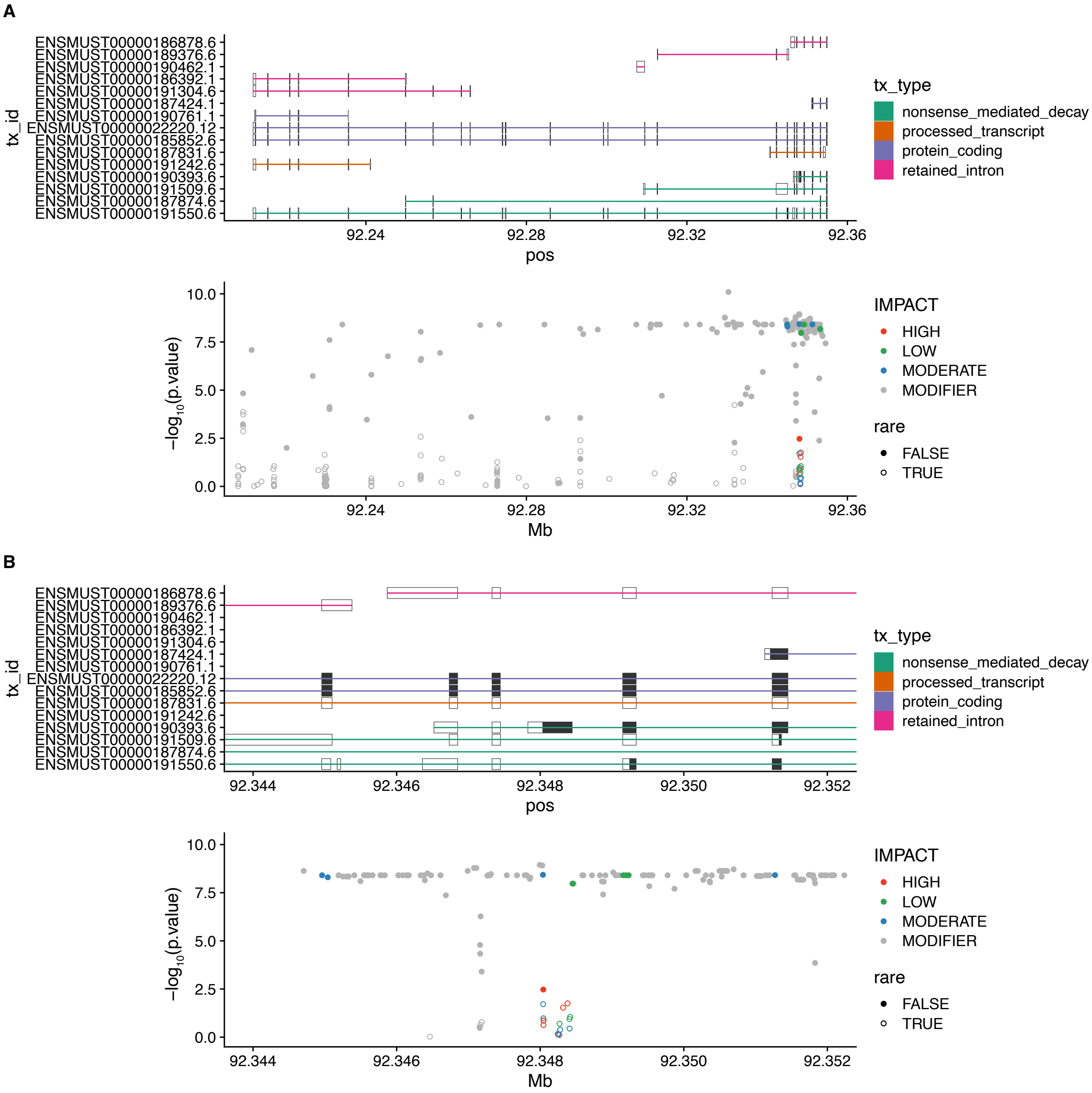


Figure 1.19 Detailed view of annotated variants within Msh3. In **A** and **B**, top panels show transcript annotations, colored by transcript type. In bottom panels, the x-axis gives the genomic coordinate and the y-axis gives the -log10 p-value of each variant for association with expansion propensity. Variants are colored by VEP-predicted impact. Filled dots represent common and empty circles represent rare variants based on the threshold identified in the previous figure. Plots are the same as those in **Fig. 1.3**, but include additional transcript annotations and rare variants. **A.** Shows the entire length of *Msh3*, whereas **B.** zooms in on the variant-dense 5’ region. High-impact rare variants overlap a 387bp LTR insertion in the “B” haplotype and likely represent variant calling artifacts.

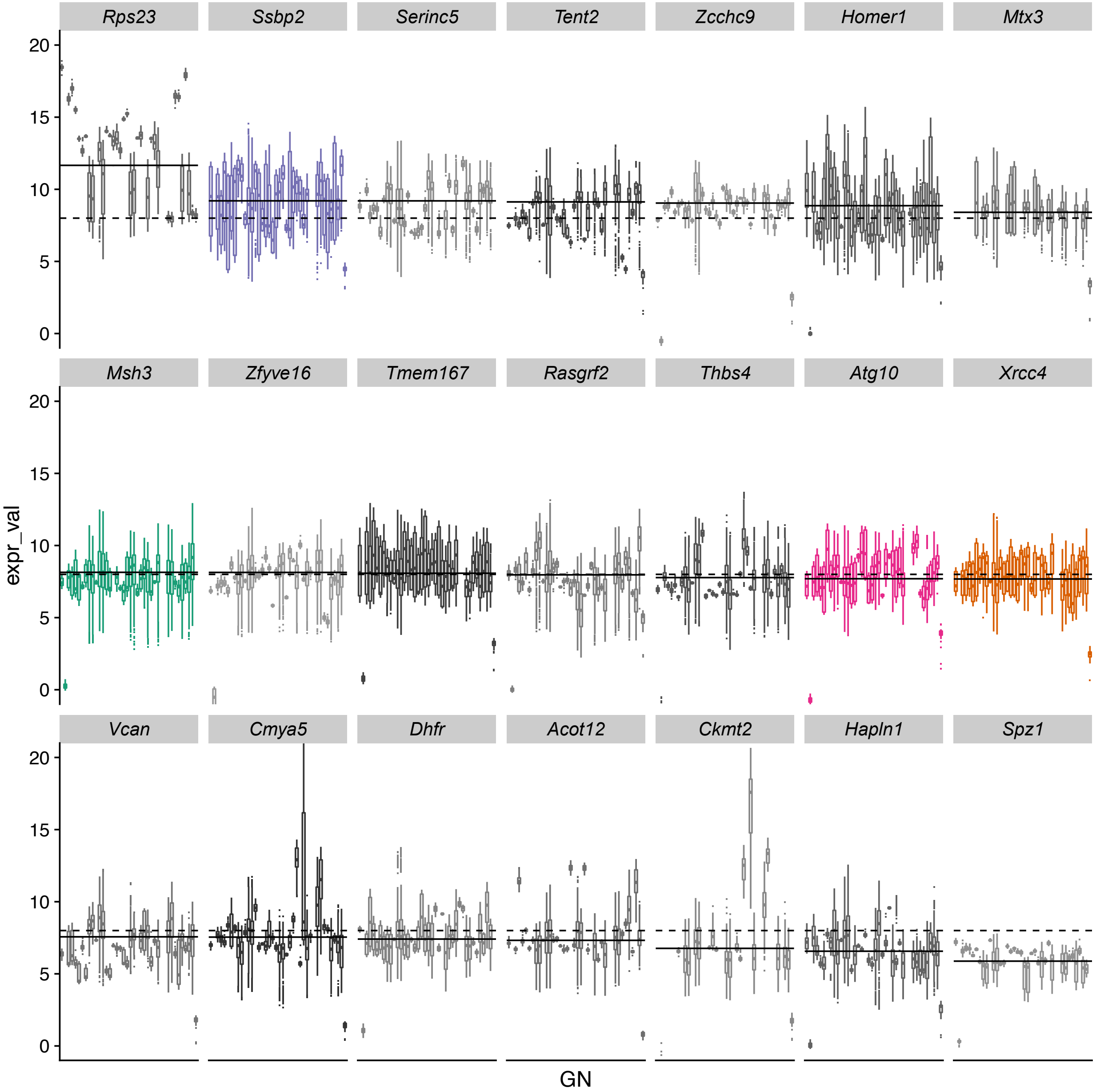


Figure 1.20 Overall gene expression levels for genes within the QTL region. Boxplots show distributions of normalized gene expression levels for each of the protein coding genes within the QTL confidence interval for tetranucleotides. Each gene is shown in a separate panel. Distributions are ordered by GeneNetwork dataset id (x-axis) and panels are ordered by the median gene expression level across all datasets (solid horizontal line). GeneNetwork datasets are normalized using a "2z+8" method (Freeman et al. 2011). The expected average value of 8 is shown as a dashed horizontal line.

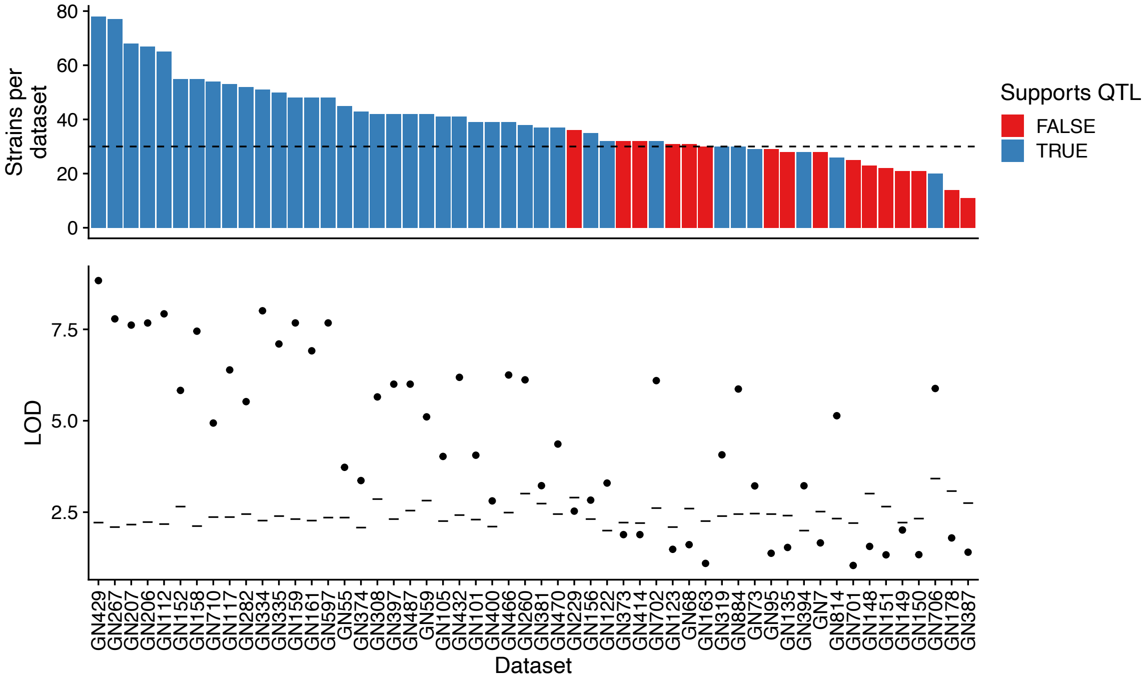


Figure 1.21 Summary of expansion propensity QTL signal detection using strains available in each gene expression dataset. The top panel shows the number of strains included in each expression dataset. Datasets are sorted in decreasing order by the number of strains per dataset. The dashed line indicates the minimum strain-per-dataset cutoff of 30 strains. We performed QTL analysis for expansion propensity using the subset of strains available in each expression dataset. The bottom panel shows peak LOD (black points) for each dataset. Gray dashes show the permutation-based significance threshold computed separately for each dataset. Blue bars in the top panel indicate the subset of strains available in that expression dataset was sufficient to reproduce the QTL for expansion propensity.

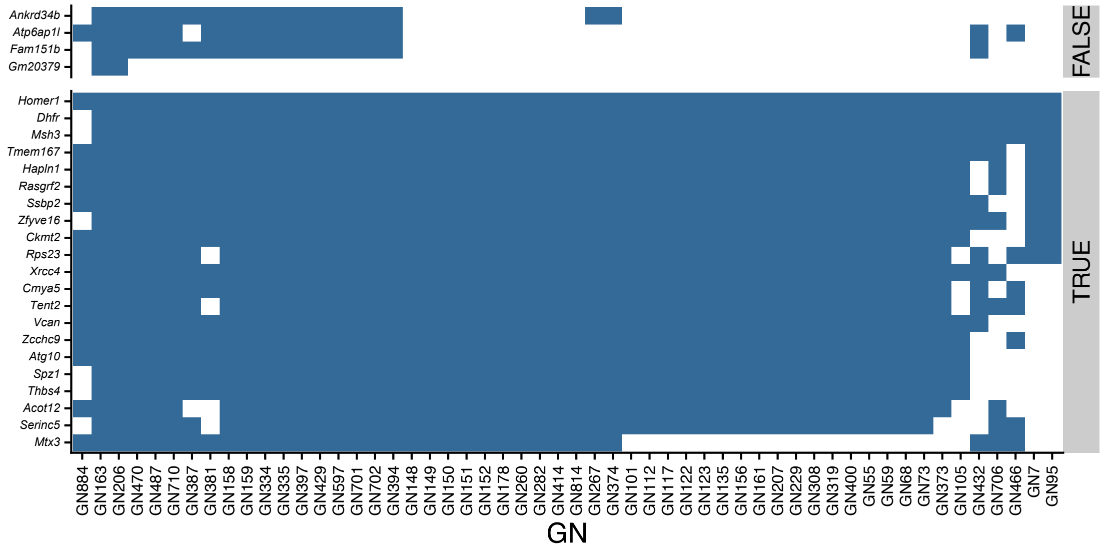


Figure 1.22 Availability of gene expression data for genes within the expansion propensity QTL. The grid indicates which protein-coding genes had gene expression values in which GeneNetwork datasets. The bottom panel shows those with values in at least 50% of the representative microarray datasets (x-axis) selected from GeneNetwork.

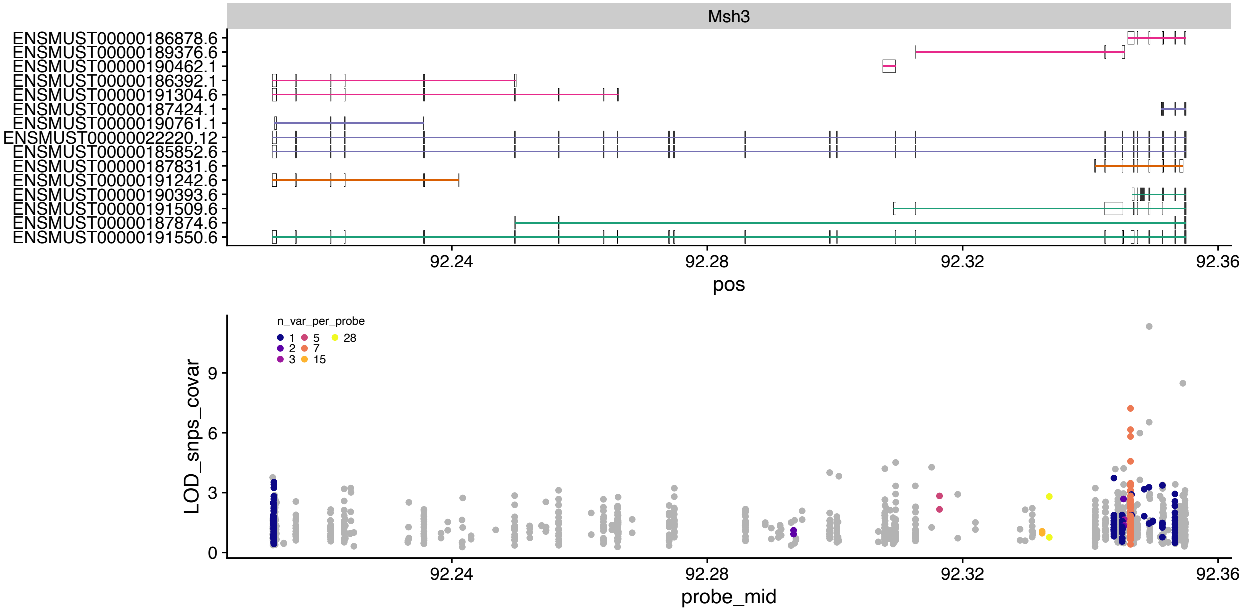


Figure 1.23 Probe-level analysis of eQTL signals at Msh3. The top panel annotates *Msh3* transcripts. In the bottom panel, each dot represents a single microarray probe. The x-axis gives the position of each probe. The y-axis gives the maximum LOD score across all datasets for each probe. Probes are colored by the number of segregating SNPs overlapping the probe coordinates. Probes not overlapping SNPs are shown in gray. Probes near the 5’ end of *Msh3* showed the strongest eQTL signals. However the majority of those overlap SNPs, which could lead to biased expression measurements and were filtered from gene expression analysis.

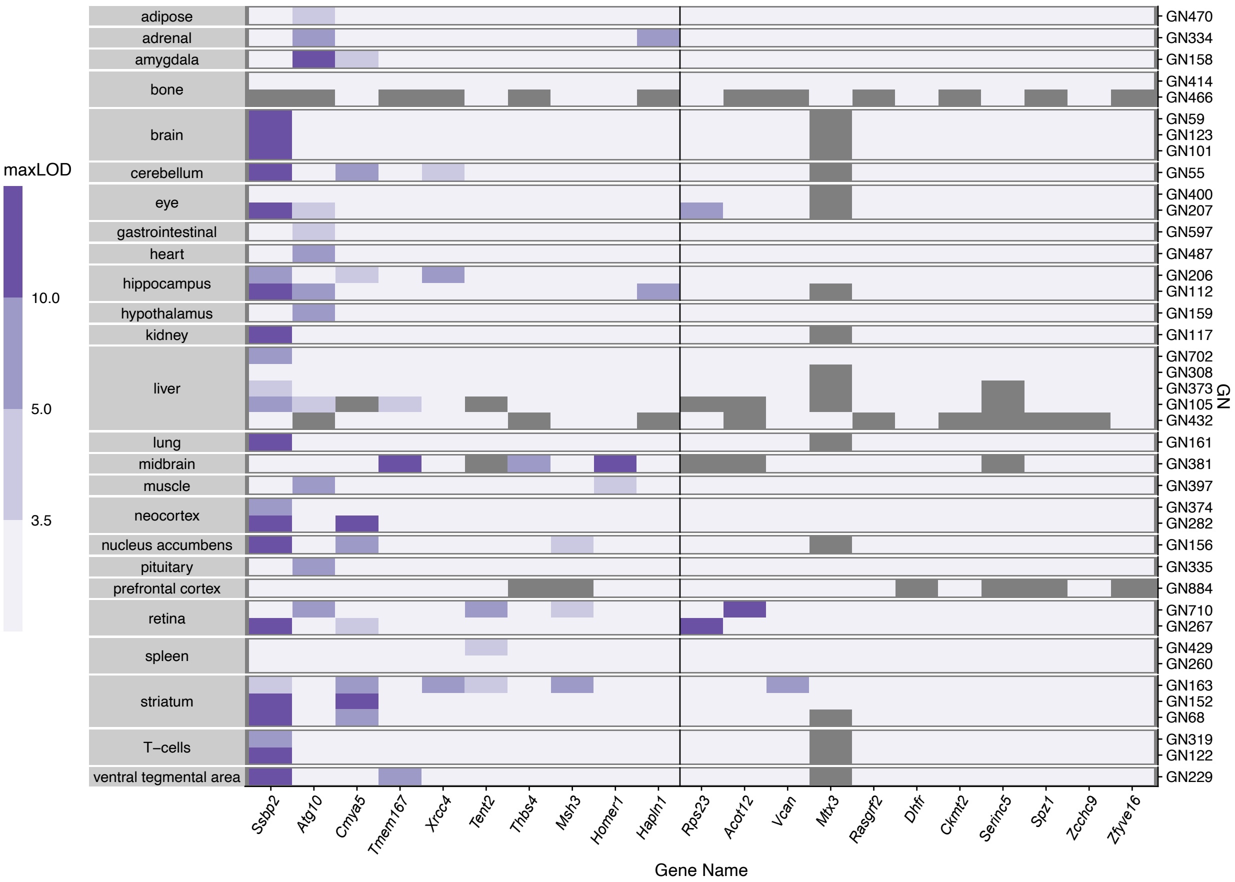


Figure 1.24 Summary of gene eQTL signals for genes contained within the QTL peak 1.5-LOD support interval for the expansion propensity phenotype. eQTL mapping was performed for each probe corresponding to a gene within the region of interest compiled across all GeneNetwork datasets. The maximum LOD value is shown for each gene (columns) in each dataset (rows, grouped by tissue). Genes are ordered from left-to-right according to the number of datasets in which the peak LOD eQTL value exceeded the permutation based threshold in that dataset. The vertical black line denotes the top 10 genes. While a single dataset is available for most tissues (primary y-axis), multiple independent datasets are available for others. GeneNetwork dataset ids are shown on the right y-axis.

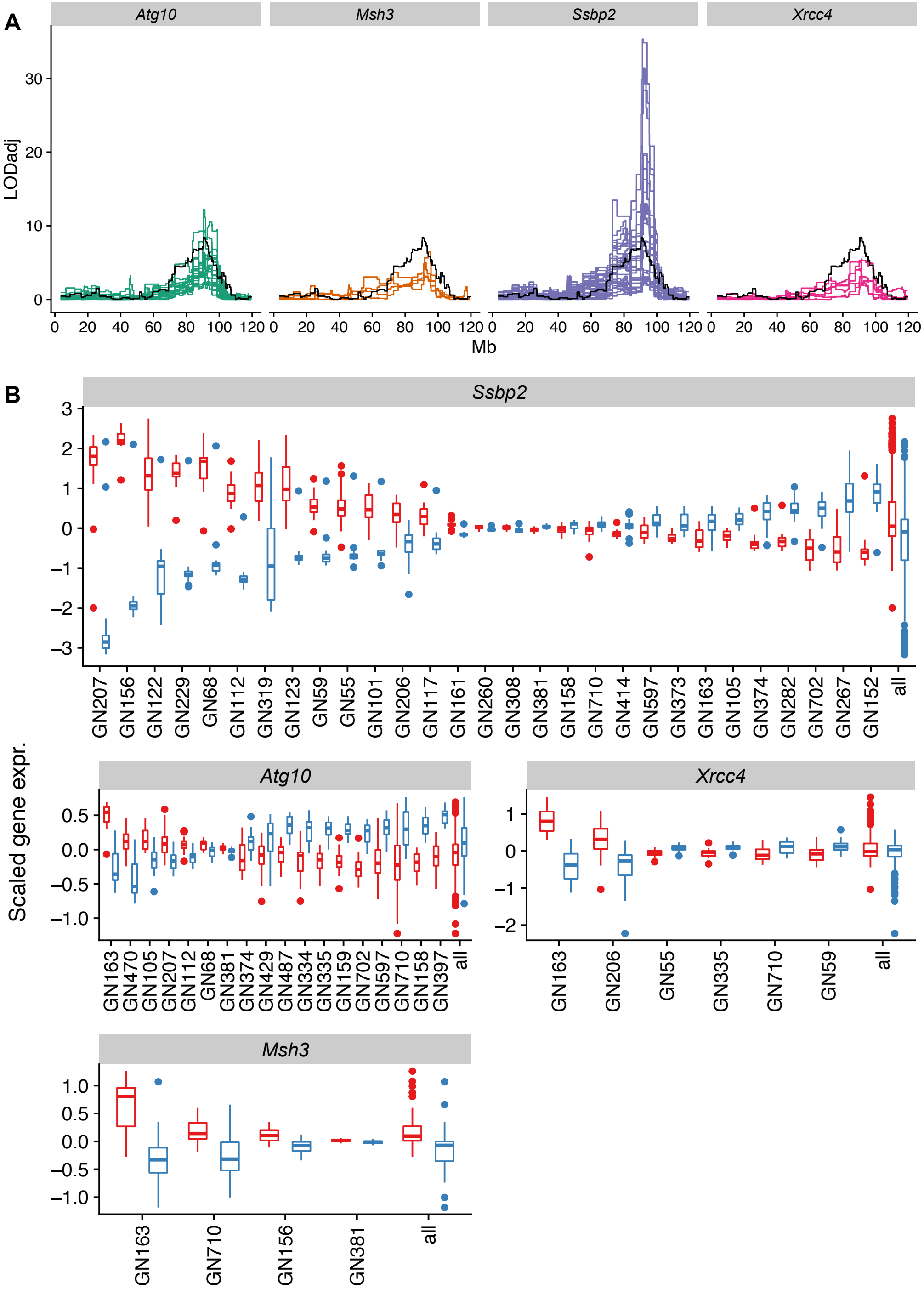


Figure 1.25 eQTL signals for DNA repair genes within the expansion propensity QTL. A. Comparison of expansion propensity and eQTL signals. LOD scores for expansion propensity are shown in black. Colored traces denote eQTL LOD scores. A separate line is shown for each expression dataset. **B. Distribution of gene expression for strains with "B" vs. "D" haplotypes.** Panels show gene expression at DNA-repair related genes for strains assigned the "B" (red) vs. "D" (blue) haplotypes at the Chr 13 expansion propensity locus. Each column denotes a different GeneNetwork expression dataset (**Supplemental Table 5**). Datasets are ranked by the difference in expression between strain groups. Only datasets where a significant eQTL was identified are shown. The far right column shows data aggregated across expression datasets.

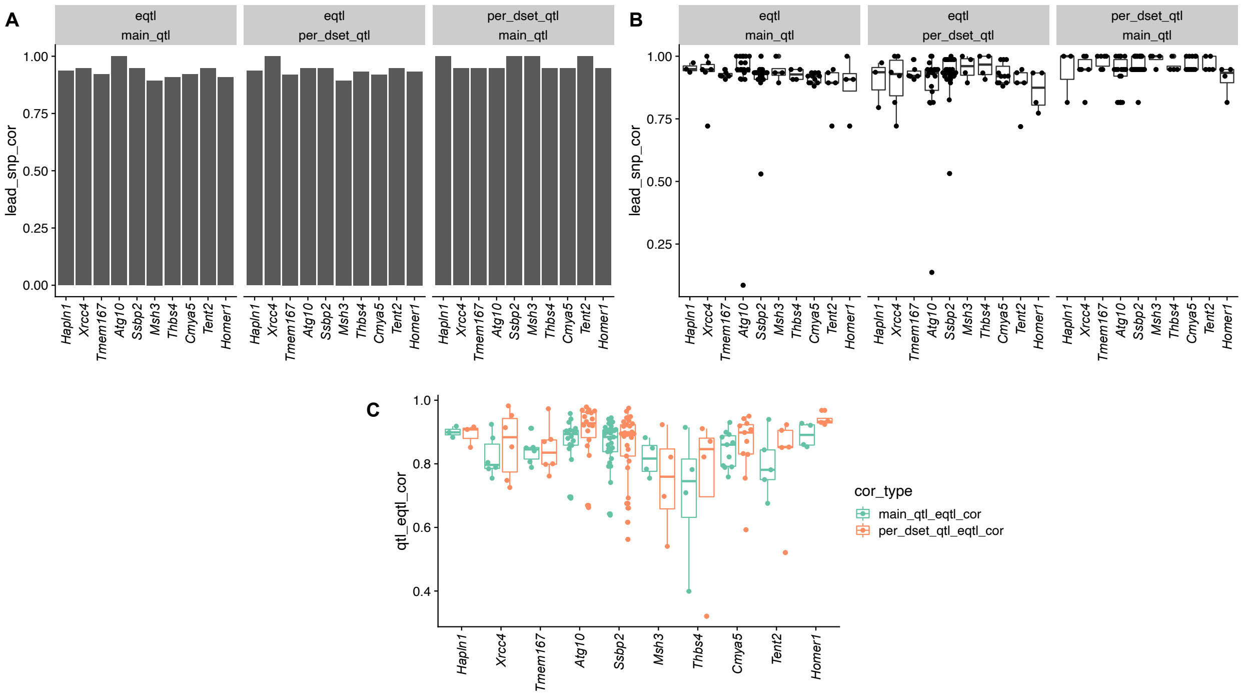


Figure 1.26 Co-localization of eQTL and expansion propensity signals. A. Correlation between the lead QTL and eQTL SNPs for each gene. For each gene, we chose the dataset with the strongest eQTL (based on LOD score) for each gene. The y-axis gives the Pearson correlation between genotypes of the lead QTL SNP and the lead eQTL SNP for each gene. **B. Co-localization across all datasets with an eQTL for each gene.** The y-axis value is the same as in A, but with a different dot for each gene expression dataset with a significant eQTL. For both **A** and **B**, in left panels the lead SNP is based on the QTL signal computed across all strains, whereas in the middle panels it is based on a QTL signal recomputed using the subset of strains available for each expression dataset. The right panels show the correlation between the lead SNP for the QTL signal based on all vs. the subset of strains. **C. Correlation of QTL and eQTL traces for each gene.** For each gene in each expression dataset with a significant eQTL, we computed the correlation between QTL and eQTL LOD scores. Green dots are computed using the main QTL signal, whereas for orange dots the QTL signal was recomputed using the subset of strains available for each expression dataset.

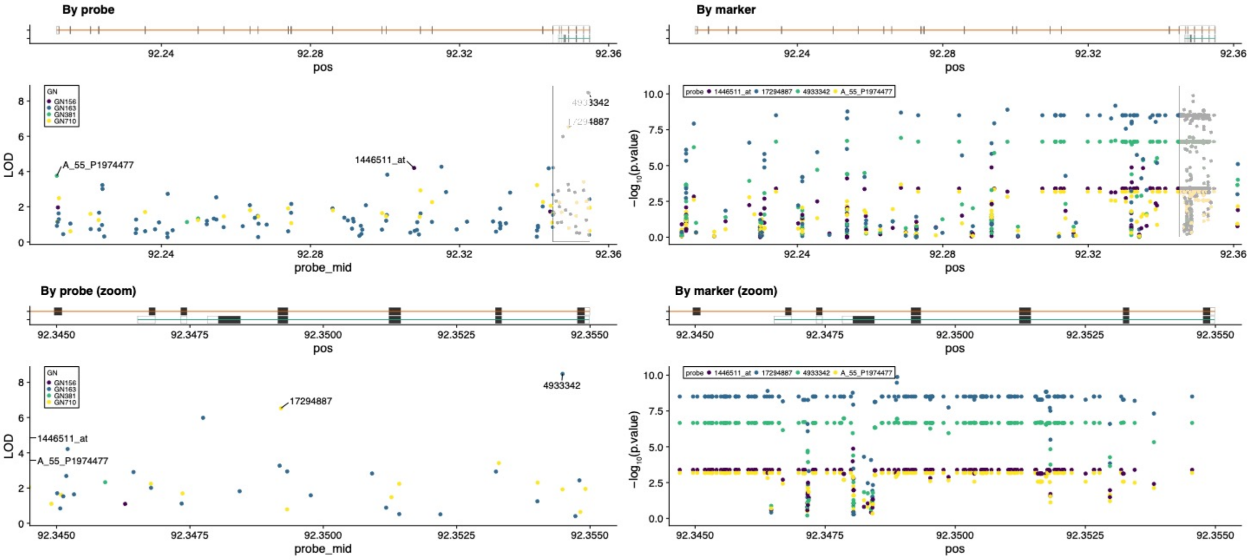


Figure 1.27 Detailed analysis of eQTL signals at Msh3. Left panels show the location of each microarray probe (x-axis) and the maximum LOD score across all variants for association with that probe. Colors represent different GN datasets. Right panels show the location of each variant (x-axis) and the best -log10 p-value across all *Msh3* probes. Colors denote different microarray probes. Bottom panels show zoomed-in views denoted by the gray rectangles in top panels, which contain both the probes and variants with the strongest eQTL signals near the 5' end of *Msh3*.

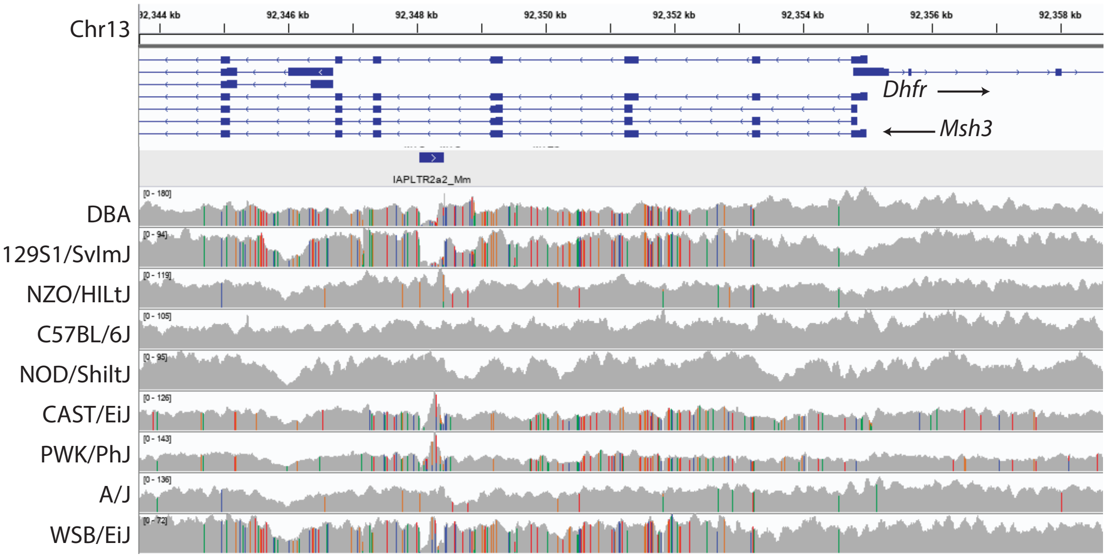


Figure 1.28 Visualization of next-generation sequencing data for classic mouse strains at the 5' end of Msh3. Top tracks show gene annotations. The middle track denotes the location of the IAP LTR element described in the main text. Bottom tracks show sequencing coverage in classic mouse strains. Colored bars indicate sequence variants compared to the mm10 reference genome, which is based on C57BL6/J. Strains 129S1/SvImJ and WSB have similar haplotypes in this region to DBA ("D"), whereas NOD is similar to C57BL6/J. Coverage profiles suggest strains DBA, 129S1/SvImJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ do not have the IAP LTR insertion present in the reference genome. The visualization was created using the Integrative Genomics Viewer (IGV).

Competing interest statement

N. Mousavi is currently employed by 10x Genomics but was not employed there when working on this study.

1.8 Acknowledgments

We thank Dr. Benjamin A. Taylor for initiating the BXD and for continued support and encouragement. We thank Drs. Gerald McClearn and Lisa Tarantino for 60 eighth-generation (G8) AI progeny from Pennsylvania State University, which contributed to epoch 3, in addition to B6D2 AI progeny (G8 and G9) to make the BXD160 through BXD186 contributed by Dr. Abraham Palmer. We thank Dr. Andrew Clark for his helpful discussion. We thank Rahel Wachs for help with illustrations. This work was supported by National Institutes of Health/National Institute on Drug Abuse (NIH/NIDA) award 1U01DA051234 (A.A.P. and M.G.). The UTHSC Center for Integrative and Translational Genomics (CITG) has supported production of the BXD colony at UTHSC and will continue to support this colony for the duration of the grant. The CITG also provides generous support for computer hardware and programming associated with GeneNetwork, as well as our Galaxy and UCSC Genome Browser instances. We thank the support of the UT Center for Integrative and Translational Genomics and funds from the UT-ORNL Governor's Chair, NIDA grants P30DA044223 and P50DA037844, and National Institute on Alcohol Abuse and Alcoholism (NIAAA) grants U01AA013499, U01AA016662, and U01AA014425. The BXD Resource at the Jackson Laboratory is supported by NIH P40OD011102 awarded to Dr. Cathleen M. Lutz. The work at Stanford University is supported by National Human Genome Research Institute grant R01HG008140. We thank The University of Tennessee Infrastructure for Scientific Applications and Advanced Computing (ISAAC) for hosting raw data and compute power. The

data sets generated and/or analyzed during the current study are available in the GeneNetwork repository (https://www .genenetwork.org/).

Chapter 1, in full, is a reprint of the material as it appears in Maksimov, M. O.*, Wu, C.*, Ashbrook, D. G., Villani, F., Colonna, V., Mousavi, N., Ma, N., Lu, L., Pritchard, J. K., Goren, A., Williams, R. W., Palmer, A. A., Gymrek, M. (2023). A novel quantitative trait locus implicates *Msh3* in the propensity for genome-wide short tandem repeat expansions in mice. *Genome Research,* 33(5): 689-702. The dissertation author was one of the two lead investigators and authors of this paper.

CHAPTER 2

xQTL: A mixture-model for identification and characterization of trans-eQTLs

2.1 Abstract

While thousands of *cis* expression quantitative trait loci (*cis*-eQTLs) have been reliably identified, consistently replicating trans-eQTL effects has proven to be challenging due to insufficient statistical power, lack of comparable tissues and cohorts, and putative false positive associations. Here, we present xQTL, a novel *trans*-eQTL method based on mixture models which infers the total number of target genes of a variant, and has improved power compared to alternative methods. We implemented xQTL and another existing *trans*-eQTL detection technique (CPMA) in an easy to use software package. Our package also includes an extensive simulation framework we developed to benchmark xQTL against existing methods. Using our simulation framework, we show that widely used correction techniques such as PCA or PEER remove effects of true *trans*-eQTLs along with technical variation. We applied xQTL on a publicly available yeast expression dataset. We identify 367 unique variants acting as *trans*-eQTLs at FDR 5% and show that the top 3 hotspots of *trans*-eQTLs are predicted to affect over half of the yeast transcriptome. xQTL analysis on RNA-sequencing from the brain hemisphere of 339 HS rats identified dozens of candidate *trans*-eQTLs. The xQTL method and simulation framework provide important resources for future *trans*-eQTL studies.

The xQTL package (including our simulation framework) can be found here: https://github.com/cynthiaewu/*trans*-eQTL

2.2 Introduction

Expression quantitative trait loci (eQTLs), or genetic variants that are associated with variation in gene expression, are thought to be major drivers of complex traits and human disease (Consortium et al. 2020). The majority of eQTL studies have focused on *cis*-eQTLs, for which the genetic variant is nearby the target gene. On the other hand, *trans*-eQTLs, for which the genetic variant and associated target gene(s) are not in close physical proximity on the genome, are thought to be important drivers of expression level variation and disease risk (Westra et al. 2013), but are far more challenging to analyze. Trans-eQTL analyses are typically severely underpowered due to the large number of putative gene-by-variant pairs resulting in a high multiple hypothesis burden. Additionally, trans effects are generally weaker than *cis* effects (Shan, 2019; Albert and Kruglyak 2015; Consortium et al. 2017) , and therefore, require larger sample sizes and stronger effects to detect compared to *cis*-eQTLs. Further, technical sources of noise, which are often unknown, can create substantial variation in expression datasets and lead to many false positive eQTL calls (Stegle et al. 2012). Finally, multiple eQTLs, both in *cis* and trans, can affect the same genes which obfuscates signals and adds to the difficulty of identifying individual *trans*-eQTLs. Thus, most identified *trans*-eQTLs have not been consistently replicated across studies due to the insufficient statistical power, lack of comparable tissues and cohorts, and potential false positive associations (Consortium et al. 2017).

Previous studies have focused on detecting *trans*-eQTLs with various methods. Albert, Bloom, et al. 2018 (Albert et al. 2018) detected *trans*-eQTLs clustering at 102 hotspot loci in yeast segregants by testing putative gene-by-variant pairs. However, in species with larger genomes, such as humans, pairwise testing is less powerful due to the large number of hypothesis tests. Another study (Kolberg et al. 2020) utilized an alternative approach that tests for association between variants and aggregate representations of expression of gene sets based on various co-expression methods. This approach identified multiple *trans*-eQTLs in blood cell types for humans that were replicated in other studies. Yet, results were highly dependent on which co-expression method was chosen for analysis. A different study (Brynedal et al. 2017) leveraged cross-phenotype meta-analysis (CPMA) (Cotsapas et al. 2011) to identify global effects of a single variant by testing if the association statistics from all genes for the variant departs from the expected distribution under the null hypothesis of no *trans* effects. One limitation of CPMA is that this approach is best suited for detecting *trans*-eQTLs influencing many genes and has low power to detect *trans*-eQTLs with a small number of target genes. Further, to our knowledge these methods have not been packaged as software tools for community use and thus remain inaccessible to the majority of researchers.

Here, we introduce **xQTL**, a novel *trans*-eQTL detection method that improves power over pairwise methods by jointly modeling the effects of an individual variant across all genes. The xQTL package additionally implements previous *trans*-eQTL detection methods for comparison. xQTL's model is similar to that employed by CPMA, but uses a more biologically plausible mixture model of effect sizes. Importantly, this enables us to infer the total number of target genes of a variant and improves power in many scenarios compared to CPMA. We additionally develop a novel open-source simulation framework to benchmark performance of xQTL against existing methods (CPMA and pairwise gene-by-variant analysis) and apply xQTL to a publicly available yeast expression dataset with 1012 meiotic segregants. We replicated 3 hotspots of *trans*-eQTLs which xQTL predicted to affect over half of the yeast transcriptome. Last, we performed xQTL analysis on RNA-sequencing from the brain hemisphere of 339 HS rats and identified dozens of candidate *trans*-eQTLs.

2.3 Results

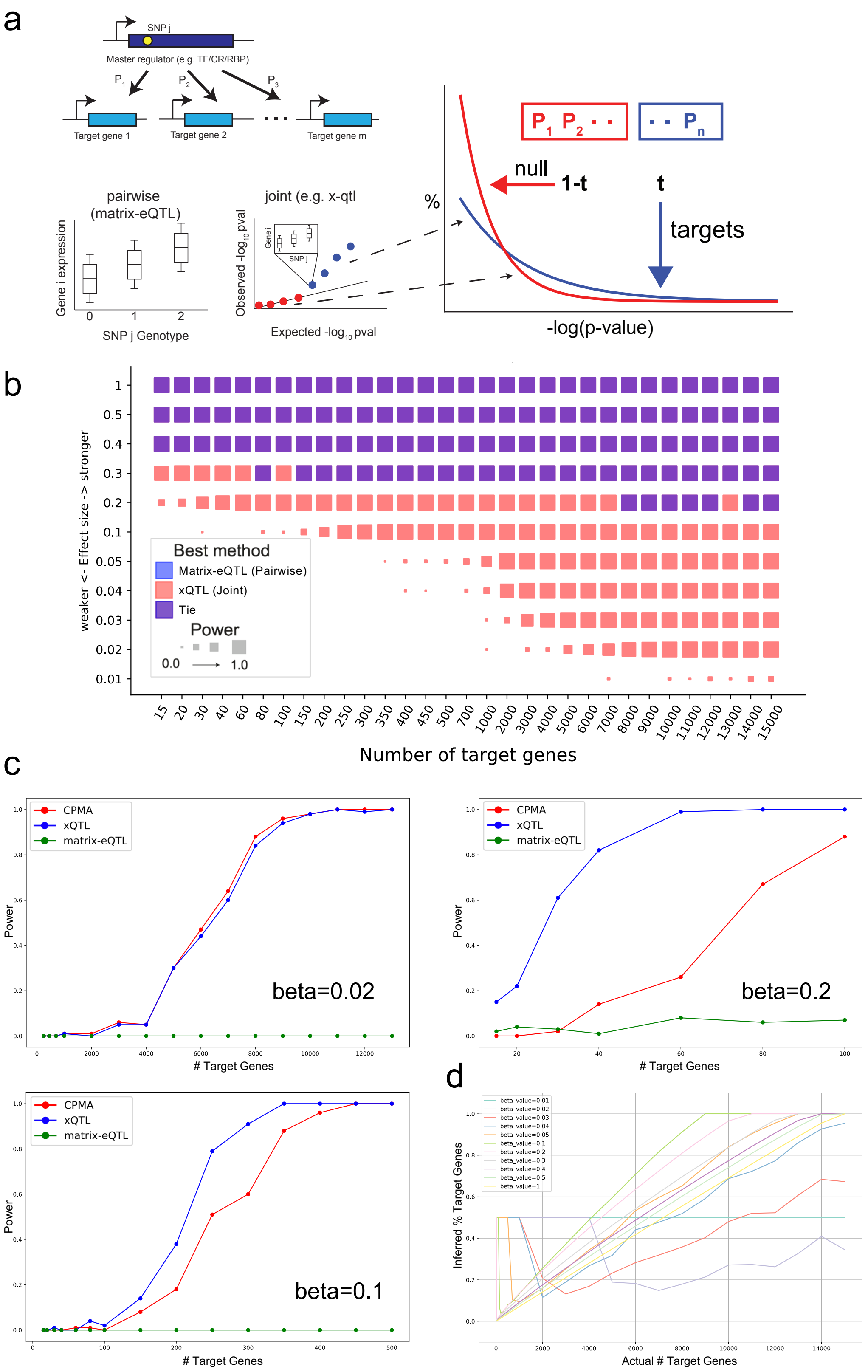
2.3.1 xQTL identifies trans-eQTLs with global effects

We developed xQTL, a method for detecting *trans*-eQTLs with effects on a large number of target genes (Fig. 2.1A). The premise of xQTL is that the p-values of association statistics for all genes of a variant of interest comes from two distinct distributions: one for the target genes and another for the non-target genes. If the variant is not a *trans*-eQTL, all -log(p-values) of association statistics will come from non-target genes, following a single exponential distribution. However, if the variant is a *trans*-eQTL, the p-values will be a mixture coming from two distinct distributions. xQTL attempts to fit this mixture distribution to learn the relative proportion of null vs. non-null (target) genes for each candidate *trans*-eQTL and outputs a likelihood ratio statistic (*Q*) which can be used to rank candidate *trans*-eQTLs and obtain the statistical significance of each candidate based on comparison to an empirical null distribution derived from permutation testing (**Methods**).

To evaluate xQTL, we developed a detailed framework to simulate expression datasets for a given sample size with a range of *trans*-eQTL effects (**Methods**). Our framework enables varying the effect size distribution and number of target genes for each *trans*-eQTL. It further models gene-gene correlation and can simulate effects of technical covariates such as those captured by PEER factors (Stegle et al. 2012). We used our framework to evaluate the power of xQTL in addition to multiple methods previously used to detect *trans*-eQTLs with various properties (Fig. 2.1B-C). We focused on two classes of methods. First, *pairwise* methods test for association between all possible SNP-phenotype pairs. If analyzing *m* SNPs and expression measurements for *n* phenotypes, pairwise methods will perform *m*n* total hypothesis tests. Second, *joint* methods consider the distribution of association statistics with all phenotypes for a particular SNP and aim to identify individual SNPs with global effects affecting a large number of target genes, rather than to identify specific SNP-gene pairs. Joint methods perform one test per SNP, and so only perform *m* total tests. xQTL falls into this second category. We additionally benchmark against the published CPMA method (Brynedal et al. 2017), which models p-values of association statistics for a particular *trans*-eQTL using a single non-null distribution, rather than as a mixture of null and non-null effects as in xQTL.

We first examined a baseline case without modeling gene-gene correlation or technical covariates. As expected, all methods show increasing power to detect *trans*-eQTLs as a function of effect size and sample size (Fig. 2.1B). For downstream analyses, we focus on results for 500 simulated individuals, similar to sample sizes for widely available expression datasets. At this sample size, naive pairwise methods are severely underpowered to detect all but the strongest effects. On the other hand, joint methods are best for detecting SNPs affecting 100 or more target genes, with at least modest effect sizes. In cases when the number of target genes and/or β effect size are very large (e.g. β>0.2 and t>100), both xQTL and CPMA are able to detect nearly all simulated *trans*-eQTLs as expected. On the other hand, our simulations show that xQTL outperforms CPMA for cases where the number of target genes and/or the effect size (β) are modest (Supplementary Fig. 2.7; Fig. 2.1C). We additionally evaluated xQTL's inferred *t* using the simulation framework. We observed that in scenarios where xQTL has sufficient power (t>~1% of target genes) the inferred *t* closely aligns with the simulated *t* for across a range of effect sizes (Fig. 2.1D). We subsequently conducted another round of simulations, modeling extensive gene-gene correlation. We found that p-values based on an empirical null

Figure 2.1 Simulation heatmap to benchmark xQTL. A. *Trans*-eQTLs can affect expression of one or more target genes. β denotes effect sizes. We focus on two classes of *trans*-eQTL detection methods. Left: pairwise methods test all possible SNP-gene pairs; Right: joint methods perform a single test for each SNP by considering all association statistics for that SNP together. Both CPMA and xQTL are joint methods and based on the sample by gene matrix which contains expression values for each gene and sample. xQTL assumes association statistics can be approximated by a mixture distribution consisting of two distributions, one for the target genes and another for the null (non-target) genes. **B.** The simulation heatmap shows the *trans*-eQTL detection method with highest power for detecting *trans*-eQTLs with varying number of target genes and effect size. The color represents the method with highest power: blue=Matrix-eQTL, red=xQTL, purple=tie between xQTL and Matrix-eQTL. xQTL has increased power over xQTL to detect *trans*-eQTLs with small number of target genes and β effect size. **C.** The power of xQTL, CPMA, and Matrix-eQTL is shown for three different effect sizes: β effect size = 0.02, 0.2, 0.1. x-QTL has the best power to detect these *trans*-eQTL except in the case of small effect sizes. **D.** The x-QTL inferred t is closely aligned with the actual simulated t for various β effect sizes except for the cases where the actual t is very small.


a
SNP j
Master regulator (e.g. TF/CR/RBP)
P1
P2
P3
Target gene 1
Target gene 2
Target gene m
pairwise
(matrix-eQTL)
Gene i expression
0
1
2
SNP j Genotype
joint (e.g. x-qtl
Observed -log10 pval
Gene i
SNP j
Expected -log10 pval
%
P1 P2 . .
. . Pn
null
1-t
t
targets
-log(p-value)
b
1
0.5
0.4
0.3
0.2
0.1
0.05
0.04
0.03
0.02
0.01
weaker <- Effect size -> stronger
Best method
Matrix-eQTL (Pairwise)
xQTL (Joint)
Tie
Power
0.0
1.0
15 20 30 40 60 80 100 150 200 250 300 350 400 450 500 700 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000
Number of target genes
c
CPMA
xQTL
matrix-eQTL
Power
beta=0.02
# Target Genes
beta=0.2
beta=0.1
d
beta_value=0.01
beta_value=0.02
beta_value=0.03
beta_value=0.04
beta_value=0.05
beta_value=0.1
beta_value=0.2
beta_value=0.3
beta_value=0.4
beta_value=0.5
beta_value=1
Inferred % Target Genes
Actual # Target Genes

were well calibrated but were inflated when using a theoretical null distribution that does not take gene-gene correlation into account. (Supplementary Fig. 2.8).

2.3.2 PEER correction removes true trans-eQTL effects

We considered the effects of technical covariates, which in real data might be introduced through unknown sources during sample preparation and/or sequencing and are often the primary source of variation in raw gene expression datasets. These unknown factors are commonly adjusted for by using PEER factor analysis (Stegle et al. 2012), or Principal Components Analysis (PCA) which identify major directions of variation that can be regressed out before downstream analyses. Failure to control for these major sources of variation can result in a large number of false positive eQTL signals. On the other hand, PEER adjustment could theoretically regress out true signals due to individual eQTLs affecting many target genes.

To evaluate the tradeoff of adjusting for PEER or PCA, we performed a round of simulations with *trans*-eQTLs of varying effects and noise. Other simulation parameters such as sample size and number of genes are based on those available in the rat dataset described below. The first set of simulations include 10 eQTLs with strong global effects (β=0.5 for 1,000 target genes out of 13,000 genes). xQTL easily detects all eQTLs when no PC adjustments are made. However, after including top expression PCs as covariates, no significant eQTLs are detected (Fig. 2.2A). Further inspection of the first 10 PCs shows they are driven by the true eQTLs, whereas PCs 11-20 capture statistical noise. For the second round of simulations, we included 20 PEER factors as technical covariates in the expression dataset in addition to 10 eQTLs and adjusted for 20 expression PCs while running xQTL. Our analysis demonstrates that xQTL could identify true strong eQTLs when performing PC adjustment only if the strength of effects of the eQTLs are sufficiently large compared to the effect sizes of the

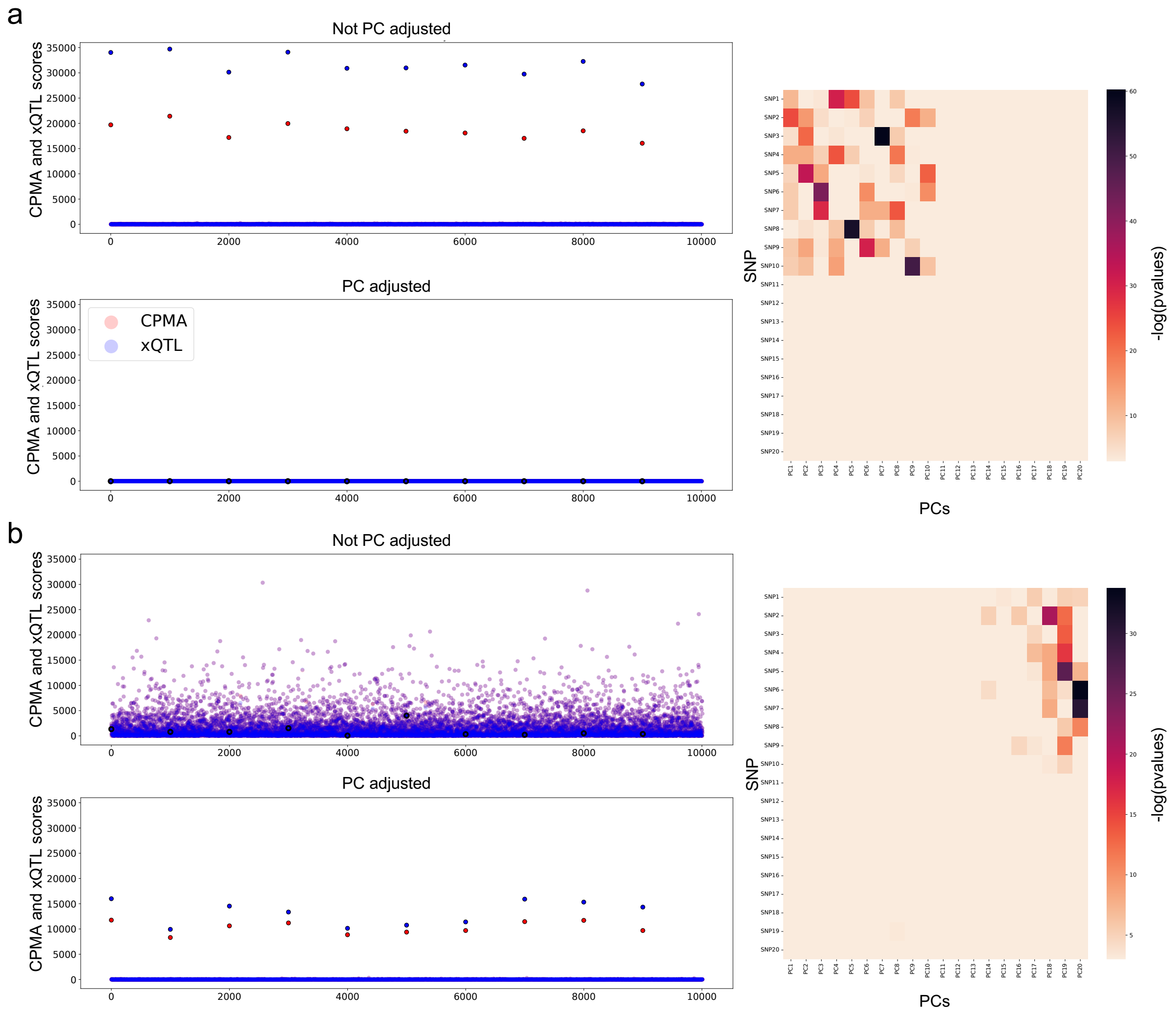


Figure 2.2 Tradeoff of *trans*-eQTL analyses with or without PC adjustment. The Manhattan plots represent various simulation and PC adjustment scenarios. The x-axis corresponds to the 10,000 simulated SNPs. The 10 simulated *trans*-eQTLs are located at each 1,000 intervals, starting from 0. The y-axis represents the CPMA and xQTL scores of the SNP. The heatmaps plot the -log(pvalue) of pairwise association testing of SNP and PCs. The x-axis represents the first 20 PCs of the expression dataset. The y-axis corresponds to 20 SNPs. SNPs 1-10 are the simulated *trans*-eQTLs and SNPs 11-20 are null SNPs. **A.** No PEER factors were simulated. *Left:* In the *trans*-eQTL analysis without PC adjustment, we can identify the 10 simulated *trans*-eQTLs. In the run with PC adjustment, we cannot identify the simulated *trans*-eQTLs. *Right:* The heatmap shows the top 10 PCs correspond to the 10 trans-eQTLs. PCs do not correspond to null SNPs. **B.** 20 PEER factors were simulated. *Left:* In the *trans*-eQTL analysis without PC adjustment, we cannot identify the simulated *trans*-eQTLs. In the run with PC adjustment, we can identify the 10 simulated *trans*-eQTLs. *Right:* The heatmap shows PCs 14-20 correspond to the 10 trans-eQTLs. PCs do not correspond to null SNPs.

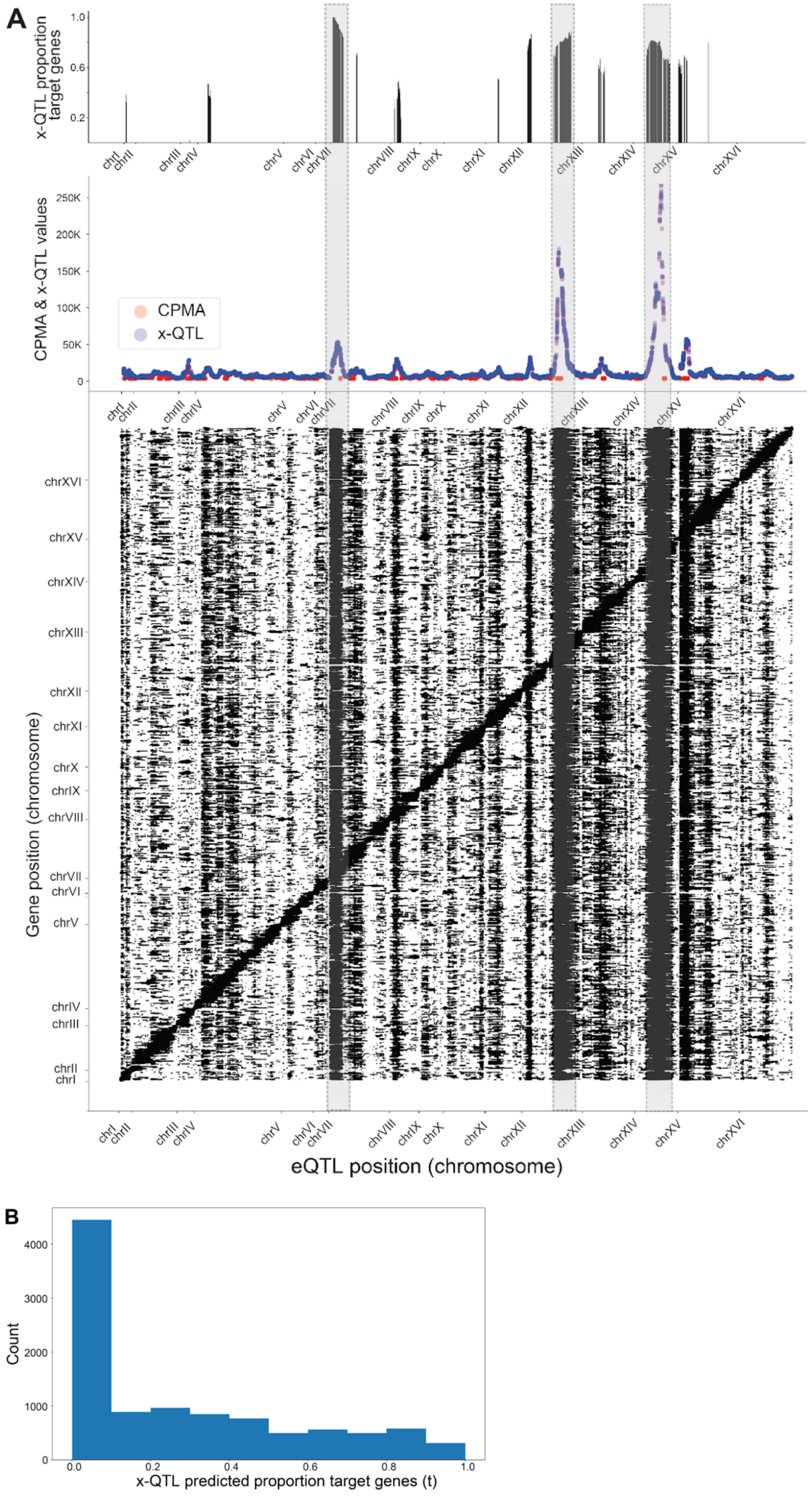
technical covariates (Fig. 2.2B). However, performing association testing separately for the PCs could identify which were driven by strong eQTLs vs. which were likely capturing technical variation. Therefore, for downstream analysis on real data we perform two parallel analyses: association testing on top expression PCs to identify major *trans*-eQTLs with global effects, and xQTL on PC-adjusted expression data to capture remaining *trans*-eQTL signals while reducing the impact of false positive signals potentially driven by technical covariates.

2.3.3 Validation of xQTL performance using yeast dataset

To evaluate xQTL on real data, we obtained a yeast expression dataset consisting of 1,012 yeast segregants for which extensive *trans*-eQTL effects had been previously identified. The dataset consists of association statistics for 5,643 genes and 11,530 variants. Visualization of the pairwise association statistics shows the expected pattern of multiple *trans*-eQTL hotspots, as was observed by the authors of the original study (Albert et al. 2018) (Fig. 2.3A, bottom).

Next, we applied CPMA and xQTL on the yeast dataset. Both are joint methods which look at the global effects of each variant across all genes. xQTL and CPMA both identify 367 unique variants acting as *trans*-eQTLs at FDR 5%. Many of these *trans*-eQTLs can be found in three eQTL hotspots on chrVII, chrXII and chrXIV. These three eQTL hotspots correspond to the thick vertical bands seen in the authors' and our eQTL map, obtained from the pairwise *trans*-eQTL method. This demonstrates that all three *trans*-eQTL detection methods were able to identify strong *trans*-eQTL hotspots, regardless of the type of tool used (Fig. 2.3A, bottom and center).

Figure 2.3: Application of xQTL, CPMA, and Matrix eQTL on yeast data. A. Bottom: - a map of Matrix eQTL values shows the genomic positions of each eQTL (x-axis) against the genomic positions of the genes whose expression they influence (y-axis). Strong diagonal band, indicating local eQTLs, and vertical bands, showing *trans*-eQTL hotspots that match the previous observation (Albert et al. 2018). Center: Map of xQTL and CPMA values shows high scoring *trans*-eQTLs are identified by both xQTL and CPMA. The *trans*-eQTL hotspots are highlighted by vertical shaded lines. Top: A map representing the xQTL proportion target genes (t) for the top 15% of xQTL scoring variants. These *trans*-eQTLs are predicted to affect over half of the transcriptome and are aligned with the trans-eQTL hotspots (shaded boxes). **B.** The histogram shows that xQTL predicts that most variants affect a small proportion (0-10%) of the genes in the yeast dataset. Note that xQTL predicts some variants to affect most of the yeast genes.


A
x-QTL proportion target genes
1.0
0.6
0.2
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
CPMA & x-QTL values
250K
200K
150K
100K
50K
0
CPMA
x-QTL
Gene position (chromosome)
eQTL position (chromosome)
B
Count
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
x-QTL predicted proportion target genes (t)

xQTL reports a predicted *t* value, the proportion of target genes a variant might have, for every variant. As expected, most predicted *t* values are small (0-10%), indicating most variants have effects on only a small number of genes (Fig. 2.3B). On the other hand, variants falling within the top 3 *trans*-eQTL hotspots were predicted to affect over half of the transcriptome (Fig. 2.3A, top). Altogether, this analysis of a dataset derived from an eukaryotic organism provides a demonstration of the ability of xQTL to capture *trans*-QTLs signals as well as the inference of the total number of genes impacted by the variant.

2.3.4 Genome-wide detection of trans-eQTLs in outbred rats

We utilized heterogeneous stock (HS) rats for trans-eQTL analysis. HS rats are derived from eight genetically diverse inbred founder strains and have been outbred for an average of 80 generations. As a result, their genomes are random mosaics of the eight founder haplotypes. Since these rats have been bred in controlled conditions, noise from environmental factors, which are prevalent in human datasets, is expected to be substantially reduced. Further, the breeding structure of this cohort has resulted in large LD blocks, reducing the total number of tests needed to perform genome-wide association testing albeit at the cost of fine-mapping precision. Overall, the relatively high rate of genetic diversity, large block size, and lack of environmental factors of HS rats makes it better powered than human cohorts of comparable size for eQTL mapping (Munro et al. 2022).

We performed xQTL analysis on RNA-sequencing from the brain hemisphere of 339 HS rats. After filtering, 13,182 protein-coding brain-expressed genes remained. In parallel, we obtained genotype calls based on whole genome-sequencing. To reduce the set of variants to those expected to either influence protein function or alter expression levels of a potential trans-regulator, we restricted the analysis to variants within exons or within +/-3kb of the

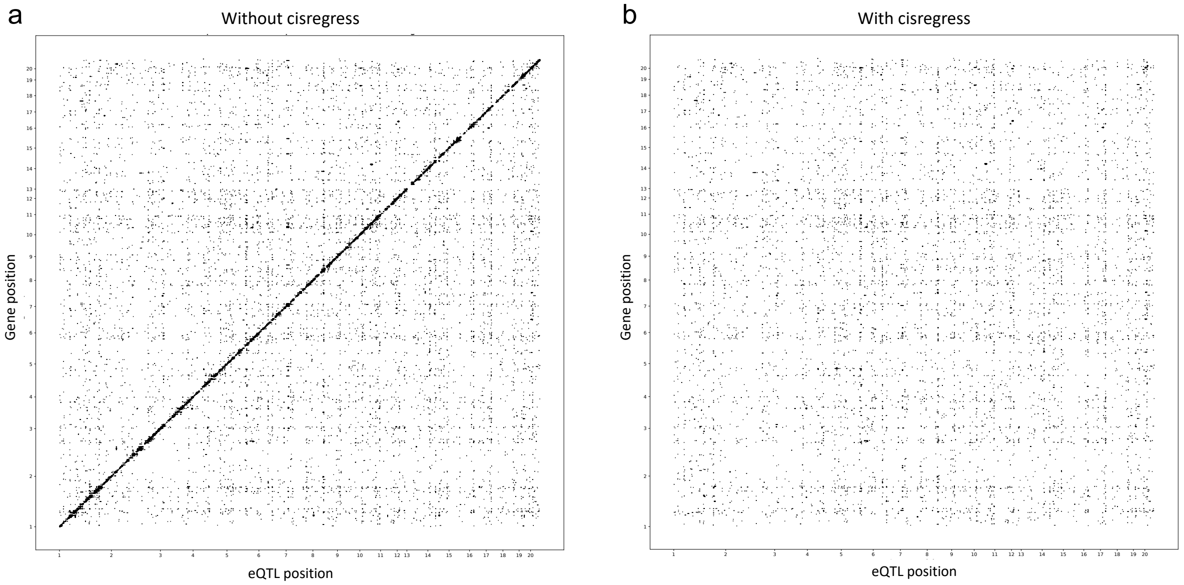


Figure 2.4 Cisregress removes the effects of *cis*-eQTLs. Plots represent a map of Matrix eQTL values showing the genomic positions of each eQTL (x-axis) against the genomic positions of the genes whose expression they influence (y-axis). The size of the dot corresponds to the strength of association of each SNP-gene pair. **A.** Matrix-eQTL was applied to the expression dataset without regressing out *cis* effects. A diagonal band representing *cis*-eQTLs is visible along with vertical *trans* bands. **B.** The effects of the top 10 *cis*-eQTLs from each gene were regressed out of the expression dataset before applying Matrix-eQTL. The diagonal band of *cis*-eQTLs is not present while the vertical bands of *trans*-eQTLs are still present.

transcription start site (TSS) of a protein coding gene. After filtering and LD-pruning, 11,002 variants remained. Further, to reduce variability in gene expression driven by *cis*-eQTLs which could impact power to detect *trans* effects, we performed *cis*-eQTL analysis with Matrix-eQTL and regressed out the effects of the top 10 *cis*-eQTLs from each gene (Methods; Fig. 2.4).

Our simulation analyses above demonstrated a tradeoff between correcting for technical variation in expression (which can remove true *trans*-eQTL signals) vs. not correcting (which can result in false positive associations) (Fig. 2.2). Therefore, we ran xQTL in two settings, with and without adjusting for expression PCs as covariates. In all cases, we adjusted for sex and genotype PCs. We observed 156 and 45 LD-independent *trans*-eQTL signals with and without controlling for expression PCs (at FDR 15%). Similar to our simulation analysis, top *trans*-eQTL signals from the analysis without adjustment corresponded to top expression PCs, (Fig. 2.5) supporting the hypothesis that expression PCs in some cases are likely capturing heritable variation.

For the run without controlling for PCs, we investigated the gene ontology (GO) processes of five top scoring *trans*-eQTL candidates identified by xQTL (Fig. 2.6). We performed a separate GO analysis for the positive and negative β target genes for each candidate. Enriched categories were different for the upregulated and downregulated categories. For each candidate, either the upregulated or downregulated category passes the FDR threshold of 0.05. As expected, the processes that passes FDR threshold are brain related. However, we observed that brain processes differ among these top-scoring variants. For example, for the chr4 candidate, the brain processes are related to morphogenesis while the brain processes are related to synaptic signaling for the chr7 candidate.

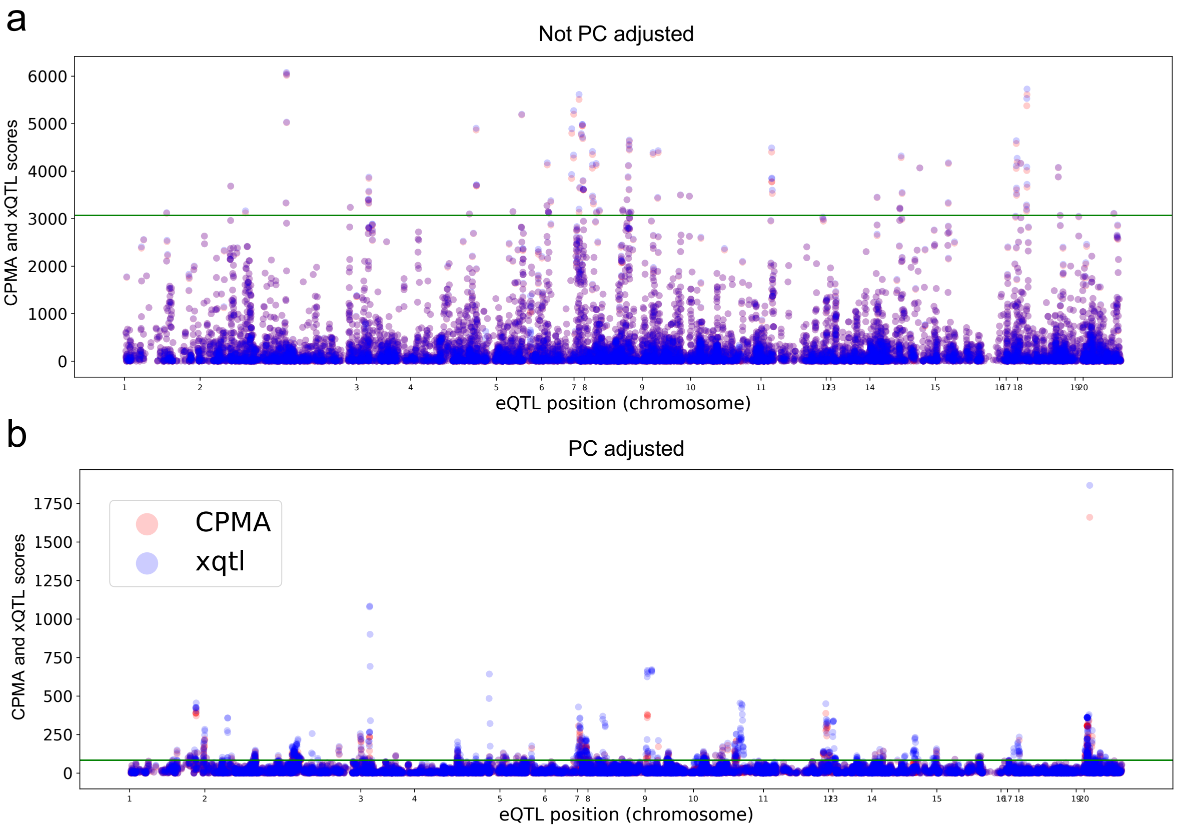


Figure 2.5 xQTL reveals *trans*-eQTL candidates. The Manhattan plots represent two different xQTL runs: with and without adjusting for expression PCs as covariates. The x-axis corresponds to the eQTL position. The y-axis represents the CPMA and xQTL scores of the SNP. The green line represents FDR of 15%. **A.** The xQTL run did not include expression PCs as covariates. We observe 45 LD-independent *trans*-eQTL signals at FDR of 15%. **B.** The xQTL run included expression PCs as covariates. We observe 156 LD-independent *trans*-eQTL signals at FDR of 15%.

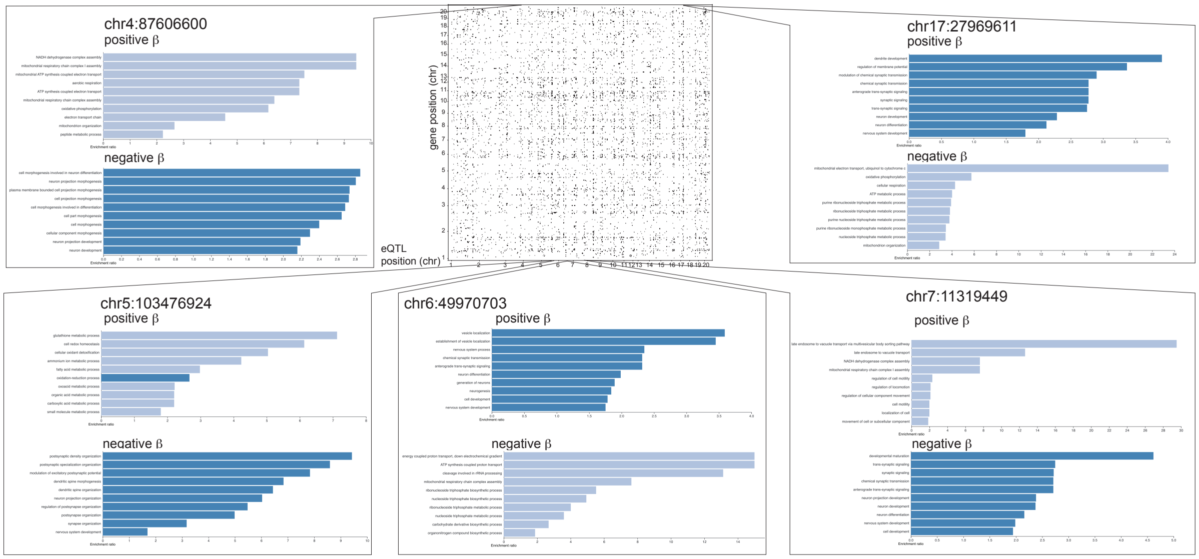


Figure 2.6 Overview of top-scoring *trans*-eQTL candidates. The middle plot represents a map of Matrix eQTL values showing the genomic positions of each eQTL (x-axis) against the genomic positions of the genes whose expression they influence (y-axis). The size of the dot corresponds to the strength of association of each SNP-gene pair. We include five top scoring *trans*-eQTL candidates identified by xQTL. For each candidate, the GO processes for the positive and negative β target genes are shown. The dark blue processes passed FDR threshold of 0.05.

2.4 Discussion

Trans-eQTL have been notoriously difficult to detect due to the many challenges of studying these variants. We developed xQTL, a novel *trans*-eQTL detection method that improves the power to detect *trans*-eQTLs over existing methods and is packaged together with a simulation framework as a publicly available toolkit. The idea behind xQTL is that the p-values of association statistics for all genes of a particular variant originate from a mixture of two distributions: genes that are targeted and ones that are not targeted by the variant. A key additional feature of xQTL is that our tool allows inference of the number of target genes for a particular *trans*-eQTL, and thus provides key biological information for these loci. This for instance can enable prioritizing the study of *trans*-acting modulators hosting a *trans*-eQTL according to the number of genes xQTL predicts are impacted by each regulator.

We additionally developed a simulation framework that allowed us to analyze various properties such as the number of target genes and β effect size of a *trans*-eQTL. The framework also enables the inclusion of the effects of gene correlation and covariates in simulated datasets. We used our simulation framework to benchmark xQTL against CPMA and observed improved power to detect *trans*-eQTL with low number of targets or small β effect size. Further, by using our simulation framework, we were able to determine the types of *trans*-eQTL each detection method is more suitable for and the *trans*-eQTLs we are still underpowered to detect. Lastly, using our simulation framework we also showed that above a threshold of minimal number of target genes, xQTL can accurately infer the number of genes targeted by a particular *trans*-eQTL.

We also analyzed the tradeoff of adjusting for PEER factors or PCs in *trans*-eQTL analyses. PEER or PCs adjustment are done to consider the effects of technical covariates that might be present in expression datasets. Failures to control for these sources of variation could results in false eQTL signals. However, on the other hand, adjusting for these factors could also remove true *trans*-eQTL effects. We demonstrate this tradeoff in our simulations and show that separate association testing for the PCs can allow us to distinguish which PCs are capturing *trans*-eQTL signals vs technical variation.

To evaluate the performance of xQTL, we used well studied *cis*- and *trans*-eQTL yeast datasets (Albert et al. 2018) and applied Matrix eQTL, CPMA, and xQTL. xQTL was able to replicate known *trans*-eQTL hotspots in this model organism and accurately estimate the number of target genes.

Last, we utilized RNA-sequencing from the brain hemisphere of 339 HS rats. We regressed out the effects of *cis*-eQTLs on the expression dataset which could obfuscate *trans*-eQTL signals. Next, we applied xQTL in two runs, with and without PC adjustment and identify dozens of *trans*-eQTL candidates.

2.5 Methods

2.5.1 Overview

We develop xQTL, a novel *trans*-eQTL detection method that helps address existing challenges of detecting *trans*-eQTLs. We first describe the regression model to detect pairwise gene-by-variant associations, then introduce the xQTL mixture model which jointly models all gene-by-variant associations for a particular variant.

2.5.2 Linear model for a single gene-by-variant pair

Here, for the sake of clarity, we consider the simplistic case where there is only a single SNP to explain the components used in the model. The model can be extended to include multiple SNPs when working with realistic datasets. We assume the following linear model to represent the association between genotype at a single SNP with expression of genes 1...m:

$$Y = X\beta + C\gamma + \varepsilon \qquad \text{(eq. 1)}$$

Where:

- Y is a $n \times m$ matrix of expression levels where Y_{ij} gives the expression of gene j in individual i. Expression values are assumed to follow a standard normal distribution.
- $X = (x^1, x^2, \ldots x^n)^T$ is an n dimensional vector of SNP genotypes for the SNP in individuals 1 through n. Genotypes for diploid organisms such as humans are encoded as 0, 1, 2, according to the count of the minor allele. For haploid organisms, genotypes are encoded as -1 (minor allele) or 1 (major allele).
- $\beta = (\beta_1, \beta_2, \ldots \beta_m)$ is an m dimensional vector of the effect sizes of the SNP on genes 1..m.
- C is a $n \times c$ matrix of covariates, where c is the number of covariates. Typical covariates include sex, age, or technical sources of variation as measured by PEER factors (Stegle et al. 2012).
- γ is a $c \times m$ matrix of effect sizes where γ_{ij} gives the effect of covariate i on gene j.
- ε is a $n \times m$ matrix of error terms and represents variation in Y not explained by genotypes or covariates. Each row of ε is modeled as a multivariate normal distribution with mean 0 and $m \times m$ variance-covariance matrix Σ. In the case of no gene-gene correlation, Σ is set to the $m \times m$ identity matrix.

2.5.3 xQTL model

xQTL analyzes the effects of an individual SNP jointly across all genes simultaneously. Under the null hypothesis that a SNP is not associated with expression of any gene, association statistics (-log p-values) of gene-by-SNP effects are expected to be exponentially distributed with $\boldsymbol{\lambda}$=1. On the other hand, if a SNP is associated with expression of a large number of genes, this distribution will depart from the null, with $\boldsymbol{\lambda}$>1.

CPMA, a related method that serves as the basis for xQTL, models association statistics using a single distribution and detects variants for which $\boldsymbol{\lambda}$!=1. This model assumes all genes are targets of the variant of interest, whereas in practice a single *trans*-eQTL likely only targets a subset of the transcriptome (Ratnapriya et al. 2019). Instead, xQTL models regression association statistics (-log p-values, denoted as a below) as a mixture of two distinct distributions (Fig. 2.1, above) corresponding to target and non-target genes:

$$\Pr\left(a = a_i \middle| t, \hat{\lambda}\right) = t\Pr\left(a = a_i \middle| \lambda = \hat{\lambda}\right) + (1 - t)Pr(a = a_i|\lambda = 1) \quad \text{(eq. 2)}$$

Where: a_i is the i$^{\text{th}}$ association statistic, *t* is the proportion of all genes that are target genes, 1/ $\hat{\lambda}$ is the mean association statistic for target genes, and $\Pr\left(a = a_i \middle| \lambda = \hat{\lambda}\right) = \hat{\lambda}e^{-\lambda\hat{a}_i}$. If the variant is not a *trans*-eQTL, *t*=0, and the above model is equivalent to the null model in CPMA.

By modeling effects as a mixture of null and non-null effects, xQTL has sensitivity to detect a broader range of *trans*-eQTLs, especially those affecting only a modest percentage of all genes in the transcriptome. Further, a key property of our model is that it can estimate the proportion of target genes (*t*) of a *trans*-eQTL, a feature not enabled by alternative frameworks including CPMA. The ability to estimate the number of target genes impacted by a *trans*-eQTL can be useful in understanding multiple biological systems and in studying molecular mechanisms of key *trans*-acting factors. For instance, some *trans*-eQTL hotspots may impact thousands of targets, whereas other *trans*-eQTLs may target only dozens of other genes (Kolberg et al. 2020). Additionally, previous reports show that the number of genomic loci bound by different transcription factors and chromatin regulators can be highly divergent (Ram et al. 2011; Garber et al. 2012). We envision that *t* will allow predicting the impact of genetic variants in *trans*-acting modulators such as transcription factors, chromatin regulators, splicing factors, or genes involved in signal transduction cascades.

2.5.4 Fitting the mixture model

xQTL fits the mixture model above to obtain maximum likelihood estimates for and *t* using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher 1994) based on the likelihood function:

$$\mathrm{Pr}\left(a_1, a_2, \dots a_m \middle| t, \hat{\lambda}\right) = \prod_{i=1}^{m} t\mathrm{Pr}\left(a = a_i \middle| \lambda = \hat{\lambda}\right) + (1-t)Pr(a = a_i|\lambda = 1) \qquad \text{(eq. 3)}$$

It then obtains a test statistic S using a likelihood ratio test:

$$S = -2ln\frac{\mathrm{Pr}(a_1, a_2, \dots a_m|0, \lambda = 1)}{\mathrm{Pr}\left(a_1, a_2, \dots a_m \middle| \hat{t}, \lambda = \hat{\lambda}\right)} \qquad \text{(eq. 4)}$$

To obtain a significance value for S, we utilize a similar method to that used by CPMA described previously (Brynedal et al. 2017). CPMA derives an empirical null distribution by simulating test statistics to recapitulate observed gene-gene correlations. The resulting null distribution is used to obtain an empirical p-value on our observed xQTL statistic. *Trans*-eQTLs are detected with this second-level significance testing (Donoho et al. 2004). The addition of the parameter *t* for xQTL increases the computational burden to obtain the test statistic. However, this is negligible compared to the resources necessary in deriving an empirical null for both CPMA and xQTL. Having an accurate empirical null distribution is highly important and gives an advantage over evaluating *t* directly from association statistics which can allow for many false positive target genes.

2.5.5 Simulation framework

Our framework simulates expression data for a set of m genes in n individuals based on a single causal SNP using the linear model above. It takes as input the sample size, number of genes, minor allele frequency (f) of the SNP, number of target genes, and effect sizes. Effect sizes can be set to a constant value for all genes, or alternatively drawn from a normal distribution with a specified standard deviation. Users may additionally input custom gene-gene correlation matrices (Σ) to model the realistic gene co-regulation patterns (default identity matrix), a custom set of covariates (default none), and values for the number of genes (default 15,000). SNP genotypes for each individual are drawn from a binomial distribution with 2 trials and probability of success f (default f=0.5). The expression value matrix Y is then simulated based on the linear model described above.

2.5.6 Simulated datasets

For our simulations, we consider the effect of a single *trans*-eQTL on *m* genes with *n* samples. Unless otherwise specified, we set m=15,000 and n=500. We vary the number of target genes and the effect size β. For non-target genes, β is set to 0. β was set to a constant non-zero value for all target genes. For each simulation, we additionally simulate data for 99 null SNPs with β=0 for all genes.

We varied the number of target genes to range from 5-15000 (ranging from 0.03 to 100% of all total genes). Effect sizes varied from 0.01 (weak effects) to 1 (strong effects). SNP minor allele frequencies were set to f=0.5. We also tested effect sizes and SNP minor allele frequencies drawn from a normal distribution and found that doing so does not have a major impact on the comparison of *trans*-eQTLs detection methods. Thus, we chose a fixed effect size and SNP minor allele frequencies to simply demonstrate the difference in power of *trans*-eQTL detection methods. For each simulated scenario of a specific *trans*-eQTL, we used 100 replicates.

To account for the extensive correlation between expression of pairs of genes observed in real datasets (Langfelder and Horvath 2008), we used scipy.stats.random_correlation to create a random gene-gene correlation matrix Σ to use in our simulations.

As covariates, we simulated technical variation due to unknown sources, based on a simulation technique previously published in the manuscript describing the PEER method (Stegle et al. 2012). We include 10 PEER factors in our simulations to demonstrate the effect technical covariates have on *trans*-eQTL detection methods. The model includes factor levels l and factor weights w for each simulated PEER factor:

$\boldsymbol{y_{ij} = \beta_j * x_i + \sum_{k=1}^{10} l_{ik} * w_{kj} + \varepsilon_{ij}}$ with i = 1, …, n samples, j = 1, …, m genes, k = 1, …, p factors

Factor levels l_{ik} for factor k were drawn from N(0, 0.6). Factor weights w_{kj} of factor k for gene j were drawn from N(0, σ^2k), where σ^2k ~ 0.8(Γ(2.5, 0.6))^2 which gives a heavy-tailed weight distribution. We utilize the PEER method (Stegle et al. 2012) to account for these factors in our datasets.

We utilized our simulation framework to evaluate the effects of adjusting for PEER factors or PCs on trans-eQTL detection. We simulated a case with no PEER factors and a case with 20 PEER factors. For both cases, our simulated datasets have 13,000 genes, 10,000 SNPs, and 345 samples. Out of the 10,000 SNPs, 10 are simulated eQTLs with effect size β=0.5 on 1,000 target genes

2.5.7 Power analysis

For each *trans*-eQTL scenario, we generated 100 simulated datasets, each with a single *trans*-eQTL. We determined the significant snps with the p-value threshold of 0.05/10,000 tests assuming 10,000 tested SNPs. In practice the p-value threshold will depend on the number of SNPs in a particular dataset. We then calculated the power to detect *trans*-eQTLs in the 100 simulated datasets for each scenario. We evaluated the power of 3 *trans*-eQTL detection methods: Matrix eQTL, CPMA, and xQTL and determined which method works best for each scenario.

2.5.8 Yeast dataset

Association statistics for yeast SNP-gene pairs were obtained from Albert, Bloom, et al. 2018 (Albert et al. 2018). This included 1012 meiotic segregants generated (Bloom et al. 2013) from a cross between the prototrophic yeast laboratory strain BY (MATa; derived from a cross between BY4716 and BY4700) and the prototrophic vineyard strain RM (MATα hoΔ::hphMX4 flo8Δ::natMX4 AMN1-BY; derived from RM11-1a). The authors obtained RNA-seq data for 5,720 genes and genotypes at 11,530 variant sites.

2.5.9 Rat dataset

RNA-sequencing from the brain hemisphere of 339 heterogeneous stock (HS) outbred rats were obtained from the RatGTeX portal (https://ratgtex.org/) The expression dataset consisted of TPM values of 32,576 genes. We filtered 398 genes in segmental duplication regions; 17,241 genes with variance > 0, variance < 50,000, IQR < 0, median > 0, and max_tpm > 2; 1,125 non-protein coding genes; 24 highly correlated genes; 606 genes with high heterozygosity, leaving us with 13,182 genes. We then quantile normalized the expression dataset. Genotype calls consisted of 6,621,609 variants. We filtered 10,144 variants in segmental duplication regions; 3,604,659 variants not in TSS +/-3kb or exons of protein coding genes; and 162,666 variants from LD pruning. We are left with 11,002 variants.

2.5.10 Trans-eQTL identification

A separate linear regression analysis was performed for each SNP–gene pair. We used Matrix eQTL (Shabalin 2012) to obtain the p-values for the association statistics for each SNP-gene pair. We then applied CPMA to obtain a CPMA statistic and xQTL to obtain a xQTL statistic for each SNP. For simulated datasets under the baseline model of no gene-gene correlation, we compared the CPMA and xQTL statistics to a X^2distribution with one degree of freedom to obtain a p-value for each SNP. Otherwise, we simulate an empirical null distribution to obtain an empirical p-value for each SNP. To obtain an empirical null distribution, we shuffled the labels of the genotype dataset. We ran Matrix eQTL with this shuffled genotype dataset and then xQTL and CPMA to get a xQTL statistic and CPMA statistic. These statistics make up the empirical null distribution due to the assumption that after shuffling genotype labels, the SNPs become null or not an eQTL. We compare the observed statistic to the empirical null statistics to obtain an empirical p-value based on how many null statistics we observe that have a more “extreme” value than the observed value. For simulated datasets that have only 100 SNPs, we used the value of 10,000 SNPs to adjust for the number of hypotheses tested with Bonferroni correction assuming 10,000 tested SNPs. As noted above, the actual number of hypotheses tested depends on the particular dataset and set of SNPs being analyzed.

Additionally, to detect *trans*-eQTLs instead of *cis*-eQTLs, we regressed out *cis* effects from the expression dataset before running *trans*-eQTL detection methods. We first run Matrix eQTL to obtain *cis*-eQTLs. We only kept *cis*-eQTLs that have less than 20 samples missing genotype data. We populated samples with missing genotype data with the most common genotype of the specific SNP for the resulting *cis*-eQTLs. Then we regressed out effects of the top 10 *cis*-eQTLs in the expression dataset with an elastic net model for each gene. We use the resulting expression dataset for *trans*-eQTL detection.

2.6 Supplementary Figures

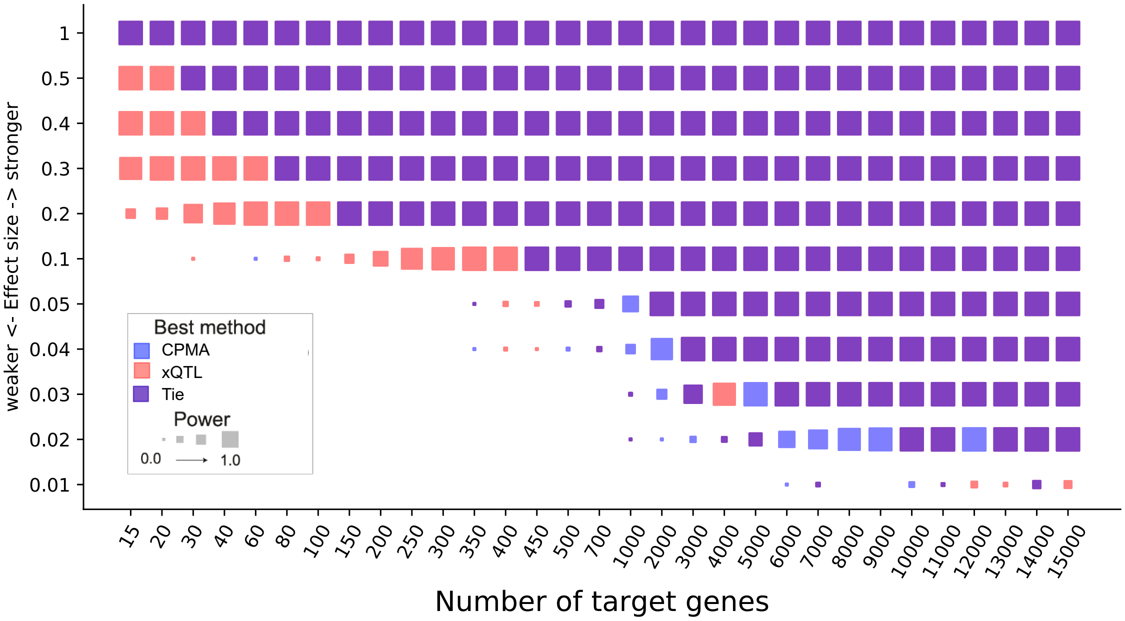


Figure 2.7 Simulation heatmap of xQTL vs CPMA. The simulation heatmap shows the method with highest power for detecting *trans*-eQTLs with varying number of target genes and effect size (assuming a baseline model with no gene-gene correlation). The color represents the method with the highest power: blue=CPMA, red=xQTL, purple=tie between xQTL and CPMA. The size of each cell of the heatmap represents the power of the corresponding best method.

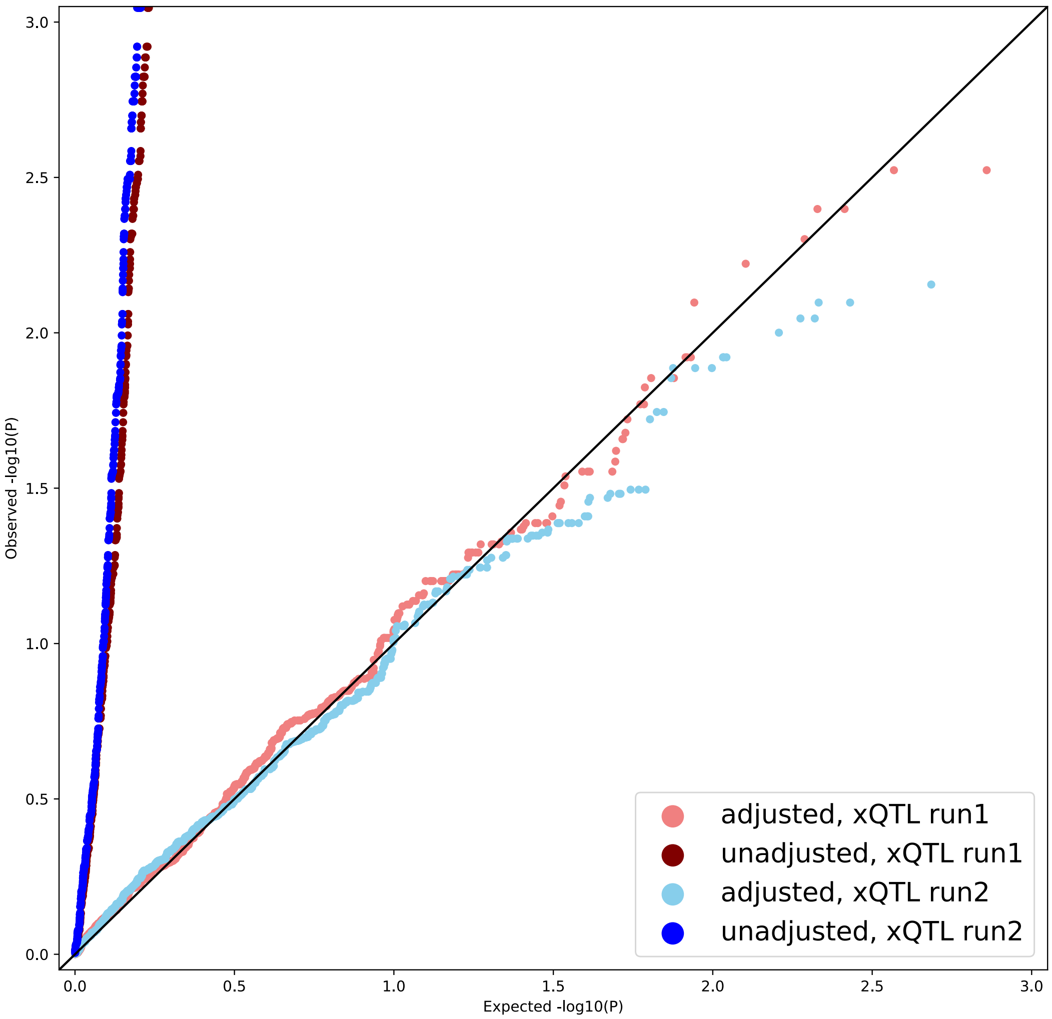


Figure 2.8 Empirical null accounts for gene correlation. QQ-plot shows -log10(p-values) of simulated null SNPs. Expression datasets are simulated with gene correlation. There are two simulations (red and blue) with different gene correlation matrices. After adjusting for gene correlation with a simulated empirical null, we observe less false positives eQTL signals.

2.7 Acknowledgments

Chapter 2, in full, is currently being prepared for submission for publication of the material by Wu, C., Xu, T., Munro, D., Mohammadi, P., Palmer, A. A., Goren, A, Gymrek, M. The dissertation author was the primary researcher and author of this paper.

CHAPTER 3

scBE-seq: A pooled, high-precision genome editing strategy with single-cell sequencing to interrogate effects of hundreds of variants

3.1 Introduction

Mutations in *trans*-acting regulators such as transcription factors (TFs), chromatin regulators (CRs), and splicing factors (SFs) can result in global transcriptomic changes leading to a variety of human diseases (Lee and Young 2013). Intriguingly, different mutations in the same gene can result in distinct phenotypes ranging from no impact to severe health consequences, often leading to completely different disease outcomes. This highlights a major challenge in human genomics: understanding the mechanistic impact of a specific mutation in its native cellular context.

A variety of computational tools have been developed for predicting the pathogenicity of a particular mutation. These primarily include gene-level constraint scores (Samocha et al. 2014; Lek et al. 2016) and machine learning based methods to predict variant-level scores (Adzhubei et al. 2013; Kircher et al. 2014; Vaser et al. 2016; Jaganathan et al. 2019). Yet, existing metrics face important limitations. (1) Gene-level information is often insufficient, as different mutations in the same gene may lead to widely different phenotypes depending on where they fall. For example, mutations in *HNF4A* may contribute to various forms of diabetes. Similarly, mutations in *MLL2* may contribute to Kabuki Syndrome, bipolar disorder, or cancer risk. Alternatively, some mutations in these genes may lead to no phenotype at all. (2) Both classes of methods do not consider tissue-specificity, and thus are not informative of what tissue is most likely to be affected by a particular mutation. (3) Despite progress in prediction methods (*e.g.*, PolyPhen2 (Jaganathan et al. 2019)), missense mutations remain challenging to interpret.

Some amino acid changes may result in dramatic changes in protein function, whereas others are benign. Still others may only affect certain cell-type specific isoforms of the gene. Thus, although a variety of prediction methods exist, they are based on computational models, have limited ground truth information for evaluation, and do not provide cell-type specific predictions of mutation impact. Predicting cell-type specific impacts of individual variants thus remains a critical unmet need in clinical variant prioritization efforts.

Pooled genome-editing assays enable interrogating effects of multiple mutations simultaneously. A variety of pooled editing approaches have been applied to assay effects of thousands of variants on expression of the edited cells (Findlay et al. 2014) or on cell survival (Hill et al. 2018). The success of these approaches relies on easily tying genotype to phenotype in each system. For example, this may be accomplished by using selection markers, gene expression based FACS soring (Gaudelli et al. 2017), or inducing the mutation of interest in the same gene or protein whose phenotype is measured (Komor et al. 2016). Yet, these techniques analyze expression of a single gene, and cannot easily be modified to measure global regulatory impacts of a single mutation.

A number of methods based on single cell technology have been used to further enhance throughput of these approaches. Notably, previous scRNAseq-based editing methods have focused on inducing deletions or gene knockouts using double stranded breaks, and could not introduce specific mutations of interest due to the low efficiency of editing using homology directed repair (HDR). Novel base editing approaches (Abid et al. 2018; Grunewald et al. 2019) now offer increased conversion efficiency of up to 50%-75% and make variant-level pooled scRNAseq approaches feasible.

Here, we introduce scBE-seq (single cell base editor sequencing), which combines a pooled, high-precision genome editing strategy with single-cell sequencing assays to simultaneously interrogate the effects of hundreds of variants. Key advantages of scBE-seq over existing methods include: (***i***) ability to induce precise mutations of interest using high efficiency base editing, rather than gene knockouts, and (***ii***) phenotype readout is global gene expression, which does not require mutation-specific specialized assays and can be applied to profile a large number of mutations of interest. Genotype-phenotype relationships for every potentially pathogenic variant in a gene can be profiled in a single experiment. Importantly, this approach can be directly extended to interrogate transcriptome-wide effects of mutations in theoretically any gene and cell type of interest.

In this Chapter, I discuss my work in the initial developments of scBE-seq and the overall goals of the assay.

3.2 Results

3.2.1 Efficient base-editing of target mutations in HEK293T cells

In the development of scBE-seq, I use the BE3 editor, which converts targeted C•G base pairs to T•A. This may be substituted with alternative editors (*e.g.*, adenine base editors (ABEs) which converts A•T to G•C) depending on the mutation of interest. BE3 is composed of a fusion protein that contains a catalytically inactivated ‘dead’ Cas9 (dCas9), a cytidine deaminase (APOBEC1) and uracil glycosylase inhibitor (UGI) to prevent base excision repair. A sgRNA specifically targets the editor to a genomic region with a modification window of 1-2 nucleotides.

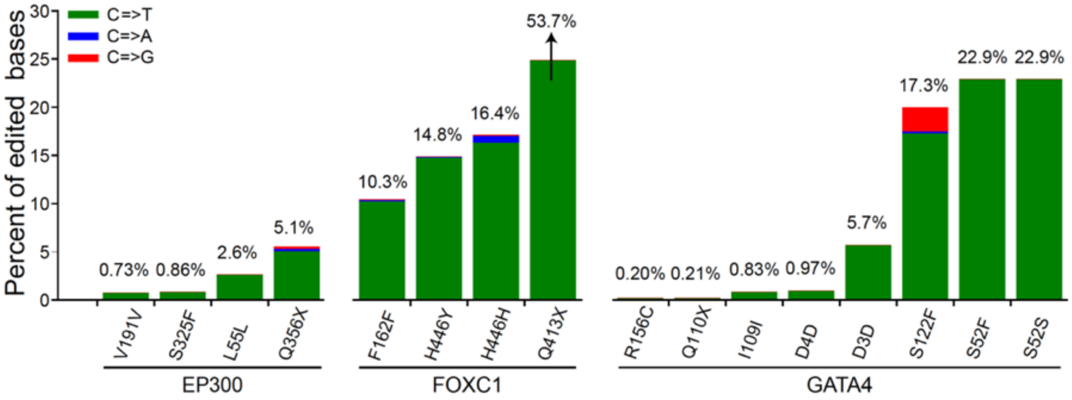


Figure 3.1: Efficiencies of successful edits using pooled base editing in HEK293. The y-axis gives the percent of each nucleotide at each position in pooled edited cells determined by targeted Illumina sequencing. **X** denotes nonsense mutation.

I employed the base editor BE3 to target 7-9 mutations in genes *EP300*, *FOXAC1,* and *GATA4*. Different BE3 editors are compatible with different protospacer-adjacent motif (PAM) sequences. For example, to generate the mutation S52F in *GATA4*, I employed VQR-BE3 (Addgene #85171) which requires PAM NGAN. To establish the Q23TER variant in *FOXC1* I used VRER-BE3 (Addgene #85173) which requires PAM NGCG. I achieved editing efficiencies of up to 53.7% (Fig. 3.1). To test for off-target RNA edits, I performed RNA-sequencing of clonal cell lines edited with *FOXC1* mutations approximately 3 weeks after transfection. I identified far fewer RNA C->T edits (~7,000) compared to previously reported results (~600,000) (Grunewald et al. 2019) for cells analyzed 24 hours after transfection (Fig. 3.2). Thus, I consider the off-target effects of the base editor to be transient and likely to have little impact on my analyses.

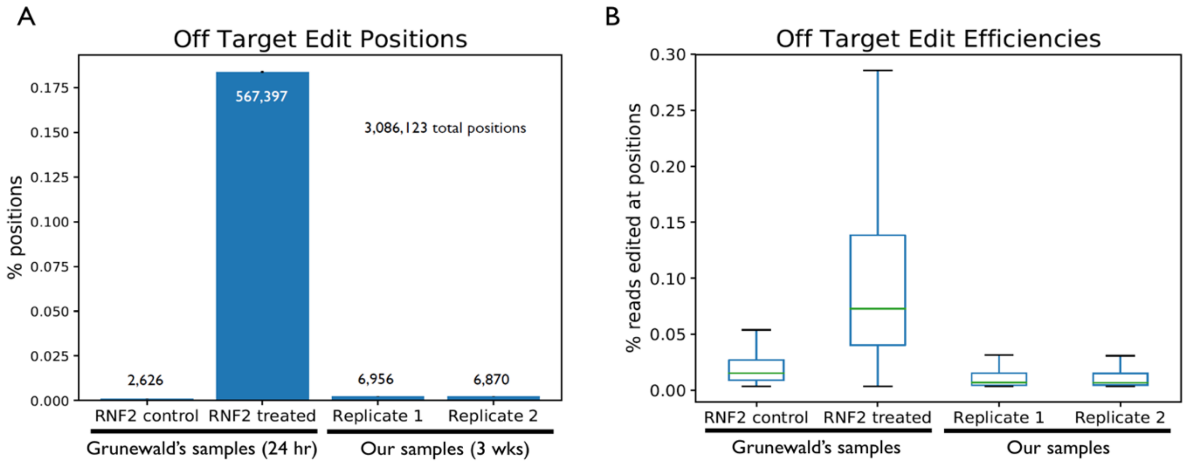


Figure 3.2: Off-target editing of RNA. Grunewald et al samples include control and treated where cells were analyzed 24hrs after transfection. We include 2 treated replicates and analyzed our samples 3 weeks after transfection. (A) The x-axis represents the samples and the y-axis is the % of positions that have off target edits. (B) The x-axis represents the samples and the y-axis is the % of reads edited at off target edit positions.

3.2.2 Types of mutations scBE-seq can interrogate.

I simulated scRNA-seq data to determine the types of mutations scBE-seq would be able to interrogate. Simulations were produced with *Splatter*, a scRNA-seq toolkit (Zappia et al. 2017). scRNA-seq data with varying cell numbers, editing efficiencies, and percentage of differentially expressed genes induced by an edit of interest, were simulated. To score how well I can distinguish between the two cell populations, I performed Louvain clustering and measured the correlation of assigned cluster labels to ground truth labels (edited vs. unedited) **(**Fig. 3.3**)**. Overall, this indicates that the requirements to study a mutation are (i) efficient genome editing of precise SNPs (>5%), and (ii) mutation has a global effect on the transcriptome (>1%). Cells with mutations with weaker effects or lower editing efficiencies cannot be distinguished from unedited cells.

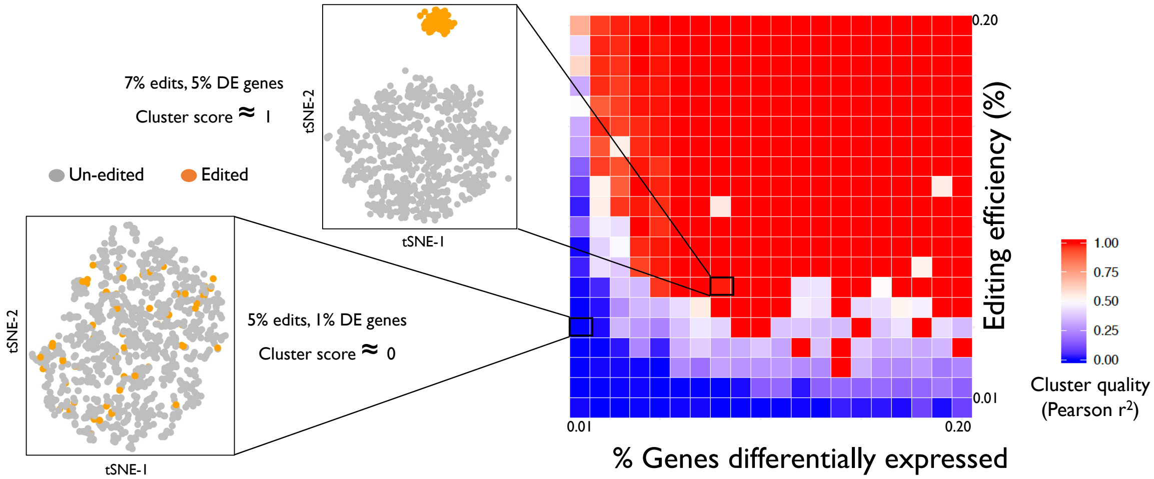


Figure 3.3: Heatmap of clusters from simulated scRNA-seq data. The x-axis represents the editing efficiency of the base editor and the y-axis is the % of genes that are differentially expressed. The color denotes the cluster quality with red denoting a good cluster. Two sections of the heatmap are zoomed in to display the clustering results from the simulated data with varying parameters and shows examples where the cluster quality is 0 and 1.

3.2.3 Script to design gRNAs

To make interrogating multiple variants scalable, I wrote a script to output all possible gRNAs in a specified gene. The script takes a gene of interest and base editor as input. It searches for all 20 bp sequences on the forward and reverse strand with a matching PAM and potential nucleotide in the edit window based on the chosen base editor. It outputs transcript, exon, and protein IDs for all possible gRNAs and reports possible "bystander" mutations that occur in the same edit window. The script has identified gRNAs that are used in the developments of scBE-seq.

3.3 Discussion

Altogether, scBE-seq involves: (1) introducing a library of mutations to proteins of interest or their regulatory elements into a cell type of interest, (2) performing single-cell RNAseq and single-cell ATAC-seq on the pool of cells, and (3) applying a custom computational pipeline to characterize genome-level effects of variants on transcriptomic and somatic mutational profiles. Ongoing developments of scBE-seq by other members of lab focus on mutations in two genes: (i) *ERCC2*, a transcription factor/DNA repair factor implicated in Xeroderma pigmentosum and (ii) *MECP2*, an X-linked chromatin regulator implicated in Rett Syndrome expected to induce widespread transcriptomic changes. Currently, developments are done with HEK293T cells, which are easily edited, and human embryonic stem cells (hESCs), which can be differentiated into multiple lineages.

Overall, we aim to use scBE-seq to perform deep mutational scanning of candidate genes. We will scale scBE-seq in male and female hESCs to interrogate all possible C->T and A->G mutations in three additional DNA repair and chromatin regulators/transcription factors, for a total of 6 genes. We target genes harboring known pathogenic variants of interest (DNA repair genes *MSH2* and *ERCC8*, and transcription/chromatin regulatory genes *MLL2* and *EP300*). The remaining two target genes will be based on candidates identified in **Chapter 2.**

3.4 Acknowledgments

Chapter 3, contains unpublished material by Wu, C., Shleizer-Burko, S., Goren, A, Gymrek, M. The dissertation author was the primary author of this chapter.

CHAPTER 4

Conclusions

In this dissertation, I presented several projects aiming to further our understanding of genetic variation in *trans*-acting factors. Variants in *trans*-acting factors can result in widespread transcriptomic changes leading to a variety of human diseases and traits. The work in this dissertation makes important contributions to the ultimate goal of precision genomics: the ability to interpret the impact of a specific mutation in a given individual.

In Chapter 1, we performed an unbiased genome-wide scan for regulators of repeat expansion propensity and identified the mismatch repair protein MSH3 as a strong *trans*-acting factor affecting germline mutation patterns in recombinant inbred mice. Varying mutation patterns due to inherited variants in and near *Msh3* are most pronounced at long tetranucleotide repeats. Importantly, we also demonstrate a potential evolutionary tradeoff in which elevated *Msh3* leads to increased repeat expansions whereas *Msh3* deficiency results in a higher rate of short insertions and deletions.

In Chapter 2, we presented xQTL, a novel *trans*-eQTL detection method based on a biologically plausible mixture model of non-target and target genes. To benchmark our tool, we also developed a simulation framework and show that xQTL has improved power over traditional *trans*-eQTL methods. We applied xQTL to whole brain RNA-sequencing data from a cohort of outbred rats and identified 45 *trans*-eQTL candidates. For example, we identified a strong candidate *trans*-eQTL locus overlapping *Neurod4*, a key neuronal transcriptional factor, which xQTL estimate to regulate thousands of target genes. Importantly, this study also highlights key technical considerations regarding treatment of technical covariates when performing *trans*-eQTL detection. Altogether, xQTL allows the detection of potential *trans* candidates which can be further characterized.

In Chapter 3, we introduced scBE-seq, a pooled genome editing assay which combines base editing and scRNA-seq to enable high-throughput variant interrogation. scBE-seq enables us to introduce specific mutations of interest which offers an opportunity to validate *trans* candidates identified by xQTL from Chapter 2. I focused on the initial developments of scBE-seq to determine and optimize base editing parameters. Ongoing progress has been made by the collaboration of multiple labs. Development of scBE-seq will allow us to interrogate the global effects of variants in numerous target genes and cell types, enabling the characterization of gene expression and mutator phenotypes.

Additionally, I contributed to An Zheng's AgentBind project which is a deep learning framework aimed to interpret sequence context for determining transcription factor binding. This study allowed the characterization of features necessary for binding of transcription factors, a major class of *trans*-acting factors. Overall, my work in this dissertation explores various approaches to further our understanding of the impact of genetic variation in *trans*-acting factors on biological functions and complex phenotypes.

References

Abid A, Zhang MJ, Bagaria VK, Zou J. 2018. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat Commun* **9**: 2134.

Adam R, Spier I, Zhao B, Kloth M, Marquez J, Hinrichsen I, Kirfel J, Tafazzoli A, Horpaopan S, Uhlhaas S et al. 2016. Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *Am J Hum Genet* **99**: 337-351.

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.

Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. 2018. Genetics of trans-regulatory variation in gene expression. *Elife* **7**.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197-212.

Ashbrook DG, Arends D, Prins P, Mulligan MK, Roy S, Williams EG, Lutz CM, Valenzuela A, Bohl CJ, Ingels JF et al. 2021. A platform for experimental precision medicine: The extended BXD mouse family. *Cell Syst* **12**: 235-247 e239.

Ashbrook DG, Sasani T, Maksimov M, Gunturkun MH, Ma N, Villani F, Ren Y, Rothschild D, Chen H, Lu L et al. 2022. Private and sub-family specific mutations of founder haplotypes in the BXD family reveal phenotypic consequences relevant to health and disease. *bioRxiv* doi:10.1101/2022.04.21.489063: 2022.2004.2021.489063.

Barrio R, Bellanne-Chantelot C, Moreno JC, Morel V, Calle H, Alonso M, Mustieles C. 2002. Nine novel mutations in maturity-onset diabetes of the young (MODY) candidate genes in 22 Spanish families. *J Clin Endocrinol Metab* **87**: 2532-2539.

Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, de Farias TM, Moretti S, Parmentier G, de Laval VR, Rosikiewicz M et al. 2021. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res* **49**: D831-D847.

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

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Bloom JS, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L. 2013. Finding the sources of missing heritability in a yeast cross. *Nature* **494**: 234-237.

Boland CR, Goel A. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**: 2073-2087 e2073.

Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. 2021. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**.

Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen S, Yandell BS, Churchill GA. 2019. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics* **211**: 495-502.

Bryda EC. 2013. The Mighty Mouse: the impact of rodents on advances in biomedical research. *Mo Med* **110**: 207-211.

Brynedal B, Choi J, Raj T, Bjornson R, Stranger BE, Neale BM, Voight BF, Cotsapas C. 2017. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am J Hum Genet* **100**: 581-591.

Campregher C, Schmid G, Ferk F, Knasmuller S, Khare V, Kortum B, Dammann K, Lang M, Scharl T, Spittler A et al. 2012. MSH3-deficiency initiates EMAST without oncogenic transformation of human colon epithelial cells. *PLoS One* **7**: e50541.

Cleaver JE, Lam ET, Revet I. 2009. Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. *Nat Rev Genet* **10**: 756-768.

Cleaver JE, Thompson LH, Richardson AS, States JC. 1999. A summary of mutations in the UV-sensitive disorders: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy. *Hum Mutat* **14**: 9-22.

Consortium GT. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318-1330.

Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida et al. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204-213.

Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J et al. 2011. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* **7**: e1002254.

Demirbag-Sarikaya S, Cakir H, Gozuacik D, Akkoc Y. 2021. Crosstalk between autophagy and DNA repair systems. *Turk J Biol* **45**: 235-252.

Donoho, D. & Jin, J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *aos* 32, 962–994

Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC, Kucherlapati R, Edelmann W, Lunetta KL, MacDonald ME et al. 2009. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol Dis* **33**: 37-47.

Duhl DM, Vrieling H, Miller KA, Wolff GL, Barsh GS. 1994. Neomorphic agouti mutations in obese yellow mice. *Nat Genet* **8**: 59-65.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**: 1184-1191.

Ekram MB, Kim J. 2014. High-throughput targeted repeat element bisulfite sequencing (HT-TREBS): genome-wide DNA methylation analysis of IAP LTR retrotransposon. *PLoS One* **9**: e101683.

Farrall M. 2004. Quantitative genetic variation: a post-modern view. *Hum Mol Genet* **13 Spec No 1**: R1-7.

Ferreira PG, Munoz-Aguirre M, Reverter F, Sa Godinho CP, Sousa A, Amadoz A, Sodaei R, Hidalgo MR, Pervouchine D, Carbonell-Caballero J et al. 2018. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun* **9**: 490.

Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**: 120-123.

Fletcher, R. 1994. An Overview of Unconstrained Optimization. *Algorithms for Continuous Optimization* 109–143

Flower M, Lomeikaite V, Ciosi M, Cumming S, Morales F, Lo K, Hensman Moss D, Jones L, Holmans P, Investigators T-H et al. 2019. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain* **142**: 1876-1886.

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. 2019. The impact of short tandem repeat variation on gene expression. *Nat Genet* **51**: 1652-1659.

Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, Robinson J, Minie B, Chevrier N, Itzhaki Z et al. 2012. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol Cell* **47**: 810-822.

Garrison E, Siren J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36**: 875-879.

Garrison E., Guarracino A., Heumos S., Villani F., Bao Z., Tattini L., Hagmann J., Vorbrugg S., Ashbrook D. G., Thorell K. et al. pggb: the PanGenome Graph Builder.

Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. 2017. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**: 464-471.

Gibson G. 2008. The environmental contribution to gene expression profiles. *Nat Rev Genet* **9**: 575-581.

Genetic Modifiers of Huntington's Disease C. 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* **162**: 516-526.

Gomes LR, Menck CFM, Leandro GS. 2017. Autophagy Roles in the Modulation of DNA Repair Pathways. *Int J Mol Sci* **18**.

Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**: 1084-1089.

Grunewald J, Zhou R, Garcia SP, Iyer S, Lareau CA, Aryee MJ, Joung JK. 2019. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**: 433-437.

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38**: 3319-3326.

Guarracino A, Mwaniki N, Marco-Sola S, Garrison E. 2021. wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm.

Gupta S, Gellert M, Yang W. 2011. Mechanism of mismatch recognition revealed by human MutSbeta bound to unpaired DNA loops. *Nat Struct Mol Biol* **19**: 72-78.

Hamilton MJ, Newbury-Ecob R, Holder-Espinasse M, Yau S, Lillis S, Hurst JA, Clement E, Reardon W, Joss S, Hobson E et al. 2016. Rubinstein-Taybi syndrome type 2: report of nine new cases that extend the phenotypic and genotypic spectrum. *Clin Dysmorphol* **25**: 135-145.

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**: 286-298.

Haugen AC, Goel A, Yamada K, Marra G, Nguyen TP, Nagasaka T, Kanazawa S, Koike J, Kikuchi Y, Zhong X et al. 2008. Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Res* **68**: 8465-8472.

Hill AJ, McFaline-Figueroa JL, Starita LM, Gasperini MJ, Matreyek KA, Packer J, Jackson D, Shendure J, Trapnell C. 2018. On the design of CRISPR-based single-cell molecular screens. *Nat Methods* **15**: 271-274.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**: D590-598.

Huang J, Kuismanen SA, Liu T, Chadwick RB, Johnson CK, Stevens MW, Richards SK, Meek JE, Gao X, Wright FA et al. 2001. MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. *Cancer Res* **61**: 1619-1623.

Huang QQ, Ritchie SC, Brozynska M, Inouye M. 2018. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res* **46**: e133.

Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, Liu W, Lin YS, Moloney C et al. 2011. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* **7**: e1002078.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB et al. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**: 535-548 e524.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-496.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* **22**: 253-259.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.

Kingwell K. 2021. Double setback for ASO trials in Huntington disease. *Nat Rev Drug Discov* **20**: 412-413.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.

Kolberg L, Kerimov N, Peterson H, Alasoo K. 2020. Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants. *Elife* **9**.

Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**: 420-424.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471-475.

Laabs BH, Klein C, Pozojevic J, Domingo A, Bruggemann N, Grutz K, Rosales RL, Jamora RD, Saranza G, Diesta CCE et al. 2021. Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism. *Nat Commun* **12**: 3216.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **175**: 598-599.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Lee TI, Young RA. 2013. Transcriptional regulation and its misregulation in disease. *Cell* **152**: 1237-1251.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.

Li GM. 2008. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**: 85-98.

Lipkin SM, Wang V, Jacoby R, Banerjee-Basu S, Baxevanis AD, Lynch HT, Elliott RM, Collins FS. 2000. MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. *Nat Genet* **24**: 27-35.

Loeb LA, Loeb KR, Anderson JP. 2003. Multiple mutations and cancer. *Proc Natl Acad Sci U S A* **100**: 776-781.

Lopez Castel A, Cleary JD, Pearson CE. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* **11**: 165-170.

Lu F, Ma Q, Xie W, Liou CL, Zhang D, Sweat ME, Jardin BD, Naya FJ, Guo Y, Cheng H et al. 2022. CMYA5 establishes cardiac dyad architecture and positioning. *Nat Commun* **13**: 2185.

Luca R, Averna M, Zalfa F, Vecchi M, Bianchi F, La Fata G, Del Nonno F, Nardacci R, Bianchi M, Nuciforo P et al. 2013. The fragile X protein binds mRNAs involved in cancer progression and modulates metastasis formation. *EMBO Mol Med* **5**: 1523-1536.

Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. 2015. Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* **15**: 181-194.

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**: 704-714.

Manley K, Shirley TL, Flaherty L, Messer A. 1999. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet* **23**: 471-473.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122.

McNulty P, Pilcher R, Ramesh R, Necuiniate R, Hughes A, Farewell D, Holmans P, Jones L, Network RIotEHsD. 2018. Reduced Cancer Incidence in Huntington's Disease: Analysis in the Registry Study. *J Huntingtons Dis* **7**: 209-222.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932-940.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**: D412-D419.

Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, Shleizer-Burko S, Lohmueller KE, Gymrek M. 2021. Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**: 246-250.

Mousavi N, Margoliash J, Pusarla N, Saini S, Yanicky R, Gymrek M. 2021. TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**: 731-733.

Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. 2019. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res* **47**: e90.

Mukherjee P, Roy S, Ghosh D, Nandi SK. 2022. Role of animal models in biomedical research: a review. *Lab Anim Res* **38**: 18.

Munro D, Wang T, Chitre AS, Polesskaya O, Ehsan N, Gao J, Gusev A, Woods LCS, Saba LM, Chen H et al. 2022. The regulatory landscape of multiple brain regions in outbred heterogeneous stock rats. *Nucleic Acids Res* **50**: 10882-10895.

Narayanan L, Fritzell JA, Baker SM, Liskay RM, Glazer PM. 1997. Elevated levels of mutation in multiple tissues of mice deficient in the DNA mismatch repair gene Pms2. *Proc Natl Acad Sci U S A* **94**: 3122-3127.

Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH. 2015. The genome as a record of environmental exposure. *Mutagenesis* **30**: 763-770.

Payseur BA, Jing P, Haasl RJ. 2011. A genomic portrait of human microsatellite variation. *Mol Biol Evol* **28**: 303-312.

Pearson CE. 2003. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. *Trends Mol Med* **9**: 490-495.

Phillips NLH, Roth TL. 2019. Animal Models and Their Contribution to Our Understanding of the Relationship Between Environments, Epigenetic Modifications, and Behavior. *Genes (Basel)* **10**.

Pierce BL, Tong L, Chen LS, Rahaman R, Argos M, Jasmine F, Roy S, Paul-Brutus R, Westra HJ, Franke L et al. 2014. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genet* **10**: e1004818.

Pinto RM, Dragileva E, Kirby A, Lloret A, Lopez E, St Claire J, Panigrahi GB, Hou C, Holloway K, Gillis T et al. 2013. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet* **9**: e1003930.

Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ. 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res* **44**: 3750-3762.

Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**: 11 12 11-34.

R Core Team. R: A Language and Environment for Statistical Computing.

Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M et al. 2011. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* **147**: 1628-1639.

Ratnapriya R, Sosina OA, Starostik MR, Kwicklis M, Kapphahn RJ, Fritsche LG, Walton A, Arvanitis M, Gieser L, Pietraszkiewicz A et al. 2019. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet* **51**: 606-610.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Roelfsema JH, White SJ, Ariyurek Y, Bartholdi D, Niedrist D, Papadia F, Bacino CA, den Dunnen JT, van Ommen GJ, Breuning MH et al. 2005. Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. *Am J Hum Genet* **76**: 572-580.

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**: 944-950.

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, Pritchard JK, Harris K. 2022. A natural mutator allele shapes mutation spectrum variation in mice. *Nature* **605**: 497-502.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353-1358.

Shan N, Wang Z, Hou L. 2019. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics* **20**: 126.

Shimosuga KI, Fukuda K, Sasaki H, Ichiyanagi K. 2017. Locus-specific hypomethylation of the mouse IAP retrotransposon is associated with transcription factor-binding sites. *Mob DNA* **8**: 20.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet* **34**: 532-544.

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**: W452-457.

Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, Kong A, Thorsteindottir U, Shin SY, Richards HB, Consortium G et al. 2011. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat Genet* **43**: 561-564.

Spitz F, Furlong EE. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**: 613-626.

Srivastava S, Avvaru AK, Sowpati DT, Mishra RK. 2019. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics* **20**: 153.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**: 500-507.

Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161-1165.

Taylor EM, Broughton BC, Botta E, Stefanini M, Sarasin A, Jaspers NG, Fawcett H, Harcourt SA, Arlett CF, Lehmann AR. 1997. Xeroderma pigmentosum and trichothiodystrophy are associated with different mutations in the XPD (ERCC2) repair/transcription gene. *Proc Natl Acad Sci U S A* **94**: 8658-8663.

Thompson PJ, Macfarlan TS, Lorincz MC. 2016. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Mol Cell* **62**: 766-776.

Tome S, Manley K, Simard JP, Clark GW, Slean MM, Swami M, Shelbourne PF, Tillier ER, Monckton DG, Messer A et al. 2013a. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS Genet* **9**: e1003280.

Tome S, Simard JP, Slean MM, Holt I, Morris GE, Wojciechowicz K, te Riele H, Pearson CE. 2013b. Tissue-specific mismatch repair protein expression: MSH3 is higher than MSH6 in multiple mouse tissues. *DNA Repair (Amst)* **12**: 46-52.

Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, Mirceta M, Mojarad BA, Yin Y, Dov A et al. 2020. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**: 80-86.

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F, Raja A, Pennacchio LA et al. 2017. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**: 710-722 e712.

Usdin K, House NC, Freudenreich CH. 2015. Repeat instability during DNA repair: Insights from model systems. *Crit Rev Biochem Mol Biol* **50**: 142-167.

van den Broek WJ, Nelen MR, Wansink DG, Coerwinkel MM, te Riele H, Groenen PJ, Wieringa B. 2002. Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum Mol Genet* **11**: 191-198.

Vandamme TF. 2014. Use of rodents as models of human diseases. *J Pharm Bioallied Sci* **6**: 2-9.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc* **11**: 1-9.

Vasicek TJ, Zeng L, Guan XJ, Zhang T, Costantini F, Tilghman SM. 1997. Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics* **147**: 777-786.

Vilar E, Gruber SB. 2010. Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol* **7**: 153-162.

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**: 1213-1216.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**: 116-117.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**: 1798-1812.

Wang Z, McSwiggin H, Newkirk SJ, Wang Y, Oliver D, Tang C, Lee S, Wang S, Yuan S, Zheng H et al. 2019. Insertion of a chimeric retrotransposon sequence in mouse Axin1 locus causes metastable kinky tail phenotype. *Mob DNA* **10**: 17.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**: 1238-1243.

Wheeler VC, Dion V. 2021. Modifiers of CAG/CTG Repeat Instability: Insights from Mammalian Models. *J Huntingtons Dis* **10**: 123-148.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**: 3350-3352.

Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. 2017. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14**: 590-592.

Yao DW, O'Connor LJ, Price AL, Gusev A. 2020. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet* **52**: 626-633.

Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* **18**, 174 (2017).

Zhou A, Zhou J, Yang L, Liu M, Li H, Xu S, Han M, Zhang J. 2008. A nuclear localized protein ZCCHC9 is expressed in cerebral cortex and suppresses the MAPK signal pathway. *J Genet Genomics* **35**: 467-472.