THE UNIVERSITY OF CHICAGO

GENOMIC TOOLS FOR ROBUST QUANTITATIVE TRAIT LOCUS DISCOVERY AND INTERPRETATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS

BY
BRYCE MYERS VAN DE GEIJN

CHICAGO, ILLINOIS
MARCH 2016

TABLE OF CONTENTS

LIST OF FIGURES v

LIST OF TABLES vi

ACKNOWLEDGMENTS vii

1 INTRODUCTION 1
1.1 Transcription factor binding sites are the words of the genome 1
1.2 Nucleosomes change the context of binding sites 2
1.3 Modifications to nucleosomes and DNA can alter regulatory states 3
1.4 eQTLs link regulatory variation to gene expression 3
1.5 Other sequencing methods for capturing regulatory information 4
1.6 Allele specific information 5
1.7 Lymphoblastoid cell lines 6
1.8 The Yoruba population 7
1.9 Dissertation overview 7

2 WASP: ALLELE-SPECIFIC SOFTWARE FOR ROBUST MOLECULAR QUANTITATIVE TRAIT LOCUS DISCOVERY 9
2.1 Abstract 9
2.2 Overview 9
2.3 Unbiased read alignment with the WASP mapping tool 10
2.3.1 Details on using WASP for mapping 11
2.3.2 Comparing WASP mapping to N-masked and personal genome mapping . 11
2.3.3 Quantifying the fraction of reads showing imbalance 12
2.3.4 Determining the effects of unknown single nucleotide variants 13
2.3.5 Assessing the effects of mapping bias on an allele-specific study 13
2.3.6 Reads filtered by WASP 14
2.3.7 Limitations of WASP mapping 15
2.3.8 Unbiased removal of amplification effects 16
2.4 Discovery of quantitative trait loci with WASP 16
2.4.1 Overview 16
2.4.2 The combined haplotype test details 19
2.4.3 The basic model 20
2.4.4 Modeling the read depths 20
2.4.5 Correcting for GC content and other effects on expected read depth 21
2.4.6 Adjusting total read depth 21
2.4.7 GC content correction 22
2.4.8 Fitting adjustment coefficients 22
2.4.9 Modeling the allelic imbalances 22
2.4.10 Correcting for incorrect genotype calls 24
2.4.11 The combined likelihood ratio test 25
2.4.12 Estimating overdispersion parameters 25

2.4.13 Beta-Negative-Binomial parameter estimation 25
2.4.14 Beta-Binomial parameter estimation 26
2.4.15 CHT calibration 26
2.4.16 Correcting for unknown covariates using principal components 28
2.5 WASP combined haplotype test performance evaluation 28
2.5.1 Identifying known European eQTLs in 69 Yoruba LCLs 29
2.5.2 Genome-wide QTL discovery in small sample sizes of ChIP-seq data . . . 32
2.5.3 Comparing CHT to other QTL mapping strategies using simulations 33
2.5.4 Simulating read depth and allele-specific counts 33
2.5.5 Comparing QTL model sensitivities 34
2.6 Testing effects of reduced allelic imbalance 37
2.6.1 Simulating sites with reduced allelic imbalance 37
2.6.2 Results 37
2.7 CHT running time 37
2.8 Combined haplotype test caveats 39
2.9 Conclusions 39

3 IDENTIFICATION OF GENETIC VARIANTS THAT AFFECT HISTONE MODIFICATIONS IN HUMAN CELLS 41
3.1 Abstract 41
3.2 Overview 41
3.3 Data generation and quality control 42
3.3.1 Samples and cell culture 42
3.3.2 ChIP-seq 42
3.3.3 Read mapping and sample validation 43
3.3.4 Heirarchical clustering to confirm data 46
3.3.5 Controlling for allelic differences in mappability 48
3.3.6 Filtering duplicate reads 48
3.3.7 Genotype imputation and phasing 48
3.3.8 RNA-seq, DNaseI-seq and MNase-seq data 49
3.4 Mapping histone modification QTLs 49
3.4.1 Details on applying the combined test genome-wide 52
3.4.2 Permutations to assess calibration 52
3.5 Evidence that histone mark and PolII QTLs are real QTLs 52
3.6 Histone modification QTLs overlap previously identified QTLs 53
3.6.1 Distance between tested SNPs and dsQTL DHSs or eQTL TSSs 53
3.7 Expression QTLs and DNase QTLs show aggregate effects on histone modifications 57
3.8 QTL changes are often coordinated across phenotypes 61
3.9 Allelic imbalance is correlated across histone marks, PolII and DNaseI 62
3.9.1 Model details 63
3.9.2 Applying the Model 64
3.9.3 Alternative method using harmonic weighted regression 64
3.10 SNPs that are dsQTL-eQTLs consistently affect enhancer and promotor modifications 65
3.10.1 Extending the CHT to multiple sites 68

3.10.2 Identifying dsQTL-eQTLs 69
3.11 Transcription factor binding consistently alters modification levels 71
3.11.1 Extending the CHT to test for transcription factor effects 72
3.11.2 Identifying transcription factor binding sites 74
3.11.3 Verifying difference in PWM score predicts allele specific occupancy . . . 74
3.12 Conclusions 77

4 TRACKING GENETIC VARIATION EFFECTS FROM CHROMATIN TO PROTEIN . 78
4.1 Abstract 78
4.2 Data processing and quality control 79
4.2.1 Sequencing data used in this study 79
4.2.2 Mapping reads 79
4.2.3 Verifying that 4su measures novel transcription rate 80
4.2.4 Peak calling and test windows for molecular traits 81
4.2.5 Standardizing data to control for read depth and GC content effects 81
4.2.6 Heirarchically clustering read counts 82
4.3 cis-QTL mapping 84
4.4 QTL effect sizes are partially buffered at the protein level 84
4.4.1 Calculating effect sizes and correlation of cis-QTLs 89
4.5 Most variation at enhancers is not linked to downstream regulation 89
4.6 Variants that affect promoter activity usually percolate to later stages 90
4.7 Many transcription QTLs are not associated with chromatin changes 91

5 DISCUSSION AND SUMMARY 93

REFERENCES 97

LIST OF FIGURES

2.1 Example of mapping bias in personalized genomes 10
2.2 WASP mapping pipeline 12
2.3 Mapping bias across methods 13
2.4 Effect of private SNPs on mapping 14
2.5 False positive rates for allelic imbalance due to mapping 15
2.6 WASP combined haplotype test pipeline 18
2.7 Read count adjustments based on read depth and GC content 23
2.8 QQ-plots of permuted CHT 27
2.9 Identifying previously known QTLs in a new dataset 30
2.10 Comparing allelic imbalances estimated by regression and WASP 31
2.11 Identifying H3K27ac QTLs in a small sample size 32
2.12 Comparing five methods for QTL identification 36
2.13 Performance of WASP with allele-specific buffering 38
2.14 Running time of CHT with difference sample sizes 39

3.1 Clustering of ChIP-seq data 47
3.2 CHT results for histone modifications and PolII 51
3.3 CDF of distances between tested SNPs and dsQTL DHSs 54
3.4 CDF of distances between tested SNPs and eQTL TSSs 55
3.5 Multiple molecular phenotypes are associated with the same genetic variants 58
3.6 Polarized effects of dsQTLs on marks at DHS regions 59
3.7 Polarized effects of eQTLs on marks around TSSs 60
3.8 Example of a QTL for multiple molecular phenotypes 61
3.9 Correlation in allelic imbalance between data types 62
3.10 Histone modification changes at dsQTLs that are also eQTLs 67
3.11 QQ-plot of dsQTL-eQTLs 71
3.12 Polymorphisms in TF binding sites affect local histone modification 73
3.13 Difference in PWM score predicts allelic imbalance 76

4.1 Logistic regression between mRNA stability and ratio of 4sU to RNA-seq reads 80
4.2 Hierarchical clustering of all sequence data 83
4.3 Effect sizes across regulatory cascade 85
4.4 Effect sizes ascertained in a different dataset 86
4.5 Percolation of QTLs 88
4.6 Franction of percolation downstream 90
4.7 Estimates of regulatory QTLs that also affect chromatin 92

LIST OF TABLES

2.1 Number of mapped reads filtered by WASP 15
2.2 Mathematical variables used by the combined haplotype test 19
2.3 Summary of QTL methods tested 34

3.1 Total sequenced, uniquely mappable and non-duplicate mapped reads for each sample. 44
3.2 Summary of significant QTL counts 50
3.3 Enrichment of histone mark and PolII QTLs near dsQTL DHSs 55
3.4 Enrichment of histone mark and PolII QTLs near eQTL TSSs 56
3.5 Genomic locations of histone mark and PolII QTLs 56

4.1 Table of all datasets processed in this study. 79
4.2 Number of PCs that maximizes the number of QTLs for each data type. 84

ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Jonathan Pritchard. I could not have asked for a better graduate experience and dynamic. Jonathan was perfectly willing to sit down for three hours straight to map out every aspect of a project, but also gave the the freedom to pursue my own questions and methods, even when they took time to develop. I also want to thank Yoav Gilad who had a great impact on my work from the beginning, but really stepped up when Jonathan moved to Stanford and I stayed in Chicago. I was also lucky enough to have the best collaborator imaginable in Graham McVicker, who showed the the ropes in project development, paper writing, and review. We were able to play off each other's strengths in statistics and computational biology to create some really elegant and powerful solutions to our biological questions. In a sense I ended up with three mentors and, though that may sound like a nightmare to many, it worked out perfectly for me. I also feel like ended up with three or four labs when I think about the fourth floor of the Cummings building where I worked. Countless members from the Stephens, Przeworski, Novembre, and Ober labs had large impacts on both my research and day to day University of Chicago experience. This is a testament to the close-knit nature of the human genetics department at Chicago and a great environment to work in. I think the biggest impact on my Chicago experience was made by my graduate student cohort in molecular biosciences. We studied together first year and we have worked together ever since. I want to thank our program administrator, Sue Levison. She was the first person that I met when I came to Chicago– if I remember correctly she immediately took me to get a shot– and has been supporting me and keeping me on track ever since. Finally, I want to thank my parents for giving me every advantage possible in reaching my goals. They are still the best teachers that I have ever had. Much of what I know about Mathematics, Statistics, and Computer Science I owe to them and, most importantly, they gave me the confidence to succeed in graduate school.

CHAPTER 1

INTRODUCTION

It has long been known that genetics has great influence over traits and, ultimately, most gene regulation can be traced back to genome sequence. Unfortunately, the process by which genome sequence is converted into regulatory instructions complex and poorly understood. Deciphering the code of the human genome would have broad ranging impacts on how we understand human traits, disease risk, and could even open the door for targeted medicine. The endeavor poses similar challenges to understanding a book that is written in an unknown language. Ideally, we would be able fully comprehend everything about the encoding of a complete human. However, an important first step is creating a dictionary to translate each DNA element and to begin understanding the regulatory syntax that pieces together these elements into gene and gene network regulatory complexes. These are underlying goals for the research programs started by Jonathan Pritchard and Yoav Gilad, in which I have been lucky enough to be involved. My work has focused on developing and applying statistical tools to use naturally occurring variation to learn the molecular function of DNA elements.

1.1 Transcription factor binding sites are the words of the genome

The central dogma of genetics states that DNA sequence is the underlying determinant of all heritable traits. DNA encodes for RNA, which encodes for protein, which defines phenotype. However, this is an extreme oversimplification as humans are made up of hundreds of cell types, each with different functions which require different proteins. This diversity begs the question: how can a DNA sequence that is identical in every cell meet the varying needs of a plethora of cell types? The answer is in gene regulation, the process by which genetic code that is identical across cell types can lead to very different gene expression patterns in each cell.

Transcription factors and their corresponding binding sites are perhaps the most basic elements of gene regulation [1]. They are proteins which interact with short DNA sequences– typically

6-12 base pairs– to either induce or repress the expression of nearby genes. Transcription factor binding sites are heavily concentrated in regions surrounding the beginning of the genes, called promoters [2]. At the promoter site, multiple factors interact together to form the pre-initiation complex, recruit RNA polymerase II, and ultimately begin transcription of the gene. Transcription factor binding sites may also be found in clusters located distal to the core promoter region [1]. Factors bound to these regions are still able to alter transcription, even if separated by hundreds of thousands of bases or more, perhaps by the looping of DNA so that the promoter and enhancers are physically close together [3] . Factors that bind at enhancers are often tissue and temporally specific, allowing for different expression programs in various cell types.

1.2 Nucleosomes change the context of binding sites

Though there is flexibility in the exact sequences, transcription factors generally have unique consensus motifs to which they bind. It should therefore be simple to know precisely where transcription factors are bound in the genome solely by evaluating the sequence. This is not the case in practice, however, as sequences which match a transcription factor's motif often remain unbound [4]. Given that transcription factor binding sites are often found in clusters, cooperative or competitive binding likely explains some of this discrepancy. Unbound sites are also likely at least in part to be due to nucleosomes, large protein complexes around which DNA is wound. Nucleosomes play a large role in the activation or repression of transcription and add an additional layer of complexity to the process [5]. Relative to most transcription factors, nucleosome complexes are quite large, binding stretches of genome spanning 146 base pairs. Their presence has been shown to block the binding of transcription factors, making it possible for nucleosomes to affect transcription by nullifying potential enhancer or promoter regions [6]. Nucleosome positioning is partially dependent on both specific sequence , as strongly bound factors such as CTCF can organize surrounding nucleosomes [7], and general base composition, as guanine and cytosine-rich stretches of DNA are often nucleosome depleted. Nucleosome occupancy is also dynamic over time as chromatin remodelers such as the SWI/SNF chromatin remodeling complexes [8] can change the landscape of the DNA. Indeed, it is believe that some transcription factors, known as ”pioneering factors”, function by displacing nucleosomes and recruiting remodelers to further open up stretches of DNA for regulation [9].

1.3 Modifications to nucleosomes and DNA can alter regulatory states

Nucleosome occupancy alone does not explain their entire effect on regulatory state. Each nucleosome is made up of eight histone subunits with amino acid chains that protrude from the complex and interact with surrounding protein [10]. These protrusions are targets for alterations, called histone modifications, which affect the properties of the nucleosome. Some modifications are markers of chromatin state. For example, acetylation of the 27th position lysine on histone 3 (H3K27ac) is enriched in regions around active enhancers [11]. It is believed that some modifications alter interactions between chromosomes leading to more compact DNA with repressed transcription. Others are thought to recruit chromatin remodelers, which move the nucleosomes and expose binding sites for other proteins. Acetylation of the histones changes the charge of the amino acids and can lead to conformational changes that affect transcription [10]. Finally, properties of the DNA itself can be modified. DNA nucleotides, particularly cytosine, can be directly methylated to influence expression. Indeed, methylation upstream of the transcription start site has been correlated with tighter nucleosome binding and therefore reduced expression. Nonetheless, the general functions and set up of histone modifications and DNA methylation are largely unknown.

1.4 eQTLs link regulatory variation to gene expression

Single nucleotide polymorphisms (SNPs) are natural variations in DNA sequence that frequently occur throughout the genome. These are relatively widespread, and if one were to compare the sequence of two copies of a human chromosome, approximately one in seventy bases would be different [12]. Moreover, about one in 1000 bases contains a polymorphism that occurs at a frequently a population [13].

Because of their high density, SNPs often overlap and interfere with regulatory elements, causing changes in the regulatory cascade. Genomicists have focused on these perturbations in order the better understand the specificities of the regulatory elements using quantitative trait locus (QTL) analysis. The earliest QTL studies identified expression QTLs (eQTLs), polymorphisms that are associated with changes in steady state mRNA levels [14]. To do so, genomicists originally collected both genotype information and mRNA level measurements from a large number of unrelated individuals using SNP arrays and gene expression probes respectively. They then looked for SNPs with genotypes that were correlated with expression levels, interpreting those with significant associations as interfering with sites that are important for gene regulation. Since DNA sequence is set at birth and is primarily not a consequence of gene regulation, we can build a strong argument that the sites identified are indeed causing the regulatory changes (or are at least linked to causal variants). In this way, natural variation in both DNA sequence and regulatory measurements are leveraged to draw causal conclusions about genetic control of regulation.

Recent developments in whole genome sequencing has made it possible to expand our set of testable SNPs to the entire genome. We can also accurately measure expression levels of all genes using RNA sequencing (RNA-seq). Here, mRNA is isolated from a sample of cells, then reverse transcribed to form cDNA. The cDNA is then sheared, amplified by polymerase chain reaction, and sequenced, yielding millions of short (20-100 base pair) sequences, often called reads, that can be aligned to the genome in order to quantify expression of known genes. The statistical frameworks for analyses have mostly remained the same. With this new technology, however, we have to information to search for QTLs amongst millions of SNPs and tens of thousands of genes genome-wide [15].

1.5 Other sequencing methods for capturing regulatory information

While the identification of eQTLs helps us better understand the DNA elements that underly gene regulation, QTLs do not tell the complete story as they do little to reveal the mechanisms underlying the regulation. A change in expression may stem from an effect on regulation at the chromatin level, with changes to elements can designate nucleosome occupancy, histone modifications, or interfere with transcription factor binding sites. Fortunately, new technologies help us capture each of these possible mechanisms. DNase is an enzyme that preferentially cuts regions of DNA that are not bound by a nucleosome. Sequencing fragments after treatment with DNase, known as DNase-seq, can be used to quantify changes in nucleosome occupancy [16]. Chromatin immuno-precipitation followed by sequencing (ChIP-seq) involves fragmenting DNA, then using an antibody to pull down DNA fragments that are bound by a transcription factor of interest or by nucleosomes marked with specific histone modifications. Sequencing these fragments can quantify these chromatin level aspects of regulation. At the RNA level, changes in steady state mRNA may be due to differences in transcription rate or in the rate of mRNA degradation. Again, new advances help us distinguish these effects. By stopping transcription with Actinomycin D [17], then sequencing at various time points, we can get a sense of the decay rates of mRNA. Moreover, changes in regulation may occur after the mRNA stage. Sequencing fragments that are bound by ribosomes (ribo-seq) can be used to quantify the translation rate of mRNA to protein [18]. Finally, mass spectrometry can be used to quantify protein levels [19]. Together, these new techniques provide tremendous opportunities to probe various stages of gene regulation.

1.6 Allele specific information

Traditional QTL studies are performed using seventy or more individuals and regressing read depth in the region of interest for each individual against the genotype for that individual. In many cases it is not feasible to collect this many samples due to cost of experiments or tissue acquisition. Moreover, when effect sizes are small, even larger samples become necessary to find statistically significant effects with regression. However, read count data contains information untapped by regression analyses, allele specific read counts. Each sample will have two copies of every chromosome. If a read does not span a polymorphism, then it is impossible to tell which chromosome it represents. However, if a read spans a SNP that is heterozygous, the sequence of the read can be used to assign it to a specific chromosome copy. If, as is common, a DNA element acts only on it's own chromosome copy, termed a cis-interaction, then QTL effects should also be seen in these allele specific counts [15]. Allele specific information is extremely powerful for detecting QTLs in small sample sizes, but comes with dangers of artefacts. A major part of my work has been on identifying and correcting these potential problems.

1.7 Lymphoblastoid cell lines

As previously discussed, acquisition of samples can be a major challenge in genomics. This is particularly difficult when studying multiple stages of gene regulation, such as expression and chromatin accessibility, simultaneously. Ideally, the same individuals would be used for every experiment, but experiments are often performed years apart as assays advance. Because of these challenges immortalized cell lines, such as lymphoblastoid cell lines (LCLs) have risen in popularity or the study of molecular genetics. LCLs are B-cells that have been treated with Epstein-Barr virus, causing them to divide indefinitely [20]. This alleviates the problem of running out of cells when trying to perform multiple experiments on the same tissue sample. Moreover, LCLs can be frozen and stored for long periods of time without adverse effects, meaning new experiments can be performed at a later date.

One of the major drawbacks to using LCLs is that, though they are created from B-cells, they do not directly represent a naturally occurring cell type. Inducing indefinite replication in cells unsurprisingly leads to broad effects on gene expression patterns and loss of cell identify. Still, QTLs identified in LCLs have shown enrichment for associations with auto-immune related diseases such as lupus (citation), so there is at least some cell-type specific information we can glean from them. More importantly for those more broadly interested in the molecular genetics, the mechanisms behind the gene regulation are preserved in LCLs, even if that particular pathways are altered.

1.8 The Yoruba population

For many of the studies in the Gilad and Pritchard labs, we use LCLs derived from Yoruba individuals, an ethnic group living in Nigeria. These individuals have been fully sequenced by the HapMap project [21], so genotypes are already available. Because all of the Yoruba individuals are from the same population, but not closely related, structure is unlikely to be an issue in studies based on these individuals. This is not a feature of all populations. If a subset of individuals are more related than the rest, also known as population structure, their genotypes will be more similar. Additionally, their phenotypes may be more similar due to environmental effects or a combination of the genetics at many loci that related individuals share. These may cause correlations between genotype and regulatory measurements not based on the locus in question and spurious identification of QTLs.

1.9 Dissertation overview

Understanding the mechanisms of gene regulation is fundamental for both evolution and disease research. The rise of genomics has made it possible to collect a huge amount of data for the study of human polymorphisms and gene regulation. With these data it is common to look for quantitative trait loci (QTLs), polymorphisms in the genome with genotypes that are correlated with a regulatory measurement, most commonly mRNA levels. However, to understand the effects of genetic variation we must look beyond QTLs for mRNA levels in a single tissue. Gene regulation may vary across tissues and polymorphisms may take effect at many stages including chromatin, transcription, translation, or degradation levels. Studying these can introduce major challenges as these experiments are often expensive and samples hard to acquire. Moreover, to fully understand gene regulation variation we must look for patterns in local sequence context to explain why some polymorphism are QTLs and why others are not. In Chapter 2, I will present WASP, a set of tools designed to (i) remove experimental artefacts from QTL studies, (ii) account for the many sources of variation in sequencing data, and (iii) maximize power to detect QTLs in small sample sizes. I will then describe, in Chapter 3, how we applied WASP to discover QTLs for four different histone modifications in the human genome. These modifications are important markers of function and chromatin state and were measured in 10 unrelated human lymphoblastoid cell lines. Even with this limited sample size, we were able to identify hundreds of QTLs. We then extended WASP to look for consistent effects across polymorphisms with similar contexts. We found that polymorphisms interrupting transcription factor binding sites consistently alter local histone modifications and that variants impacting chromatin at distal regulatory sites frequently also direct changes in chromatin and gene expression at associated promoters. In Chapter 4, I will present the most comprehensive QTL identification in the various stages of gene regulation to date. We identified QTLs for histone modifications, chromatin accessibility, transcription, mRNA, translation, and protein levels. We then tracked effect sizes through the regulatory cascade, from chromatin to RNA to protein. We found a general consistency of effects, but buffering at the protein level. Finally, we found that many changes in enhancer activity cannot be linked to gene expression, but that once a promoter effect is identified, QTLs effects are likely to carry through to the protein level. Finally, in Chapter 5 I will discuss the main conclusions of thy graduate research and discuss future directions for further study.

CHAPTER 2

WASP: ALLELE-SPECIFIC SOFTWARE FOR ROBUST MOLECULAR QUANTITATIVE TRAIT LOCUS DISCOVERY

2.1 Abstract

Allele-specific sequencing reads provide a powerful signal for identifying molecular quantitative trait loci (QTLs), however they are challenging to analyze and prone to technical artefacts. Here we describe WASP, a suite of tools for unbiased allele-specific read mapping and discovery of molecular QTLs. Using simulated reads, RNA-seq reads and ChIP-seq reads, we demonstrate that WASP has a low error rate and is far more powerful than existing QTL mapping approaches.

Contribution

This work was done in collaboration with Graham McVicker. I did much of the development of the WASP mapping pipeline as well as the statistics for the Combined Haplotype Test. In particular, I made Figures 2.1, 2.3, 2.5, 2.7, 2.8, 2.9, 2.11, 2.12, 2.13.

2.2 Overview

Next generation sequencing data can be used to identify allele-specific signals because reads that overlap heterozygous sites can be assigned to one chromosome or the other. Molecular QTLs are associated with allelic imbalance[22, 23, 15, 24], and thus allele-specific reads can potentially augment the power of statistical tests for QTL discovery[25, 26]. However, use of allele-specific reads can introduce artefacts into many stages of analysis. Uncorrected mapping of allele-specific reads can be highly biased and can easily yield false signals of allelic imbalance [27, 28]. Homozygous sites which are incorrectly called as heterozygous are another source of false positives, and allele-specific read counts are overdispersed compared to the theoretical expectation of a binomial distribution [29]. Here we describe a suite of tools called WASP that is designed to overcome these

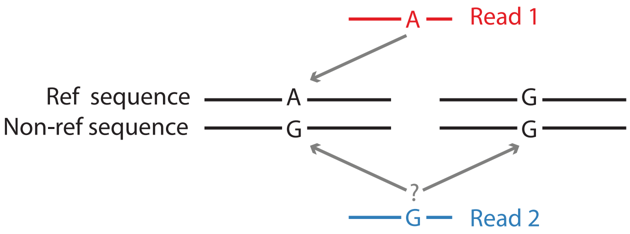

Figure 2.1: Mapping to personalized genomes can result in allelic bias because reads from one allele may not map uniquely. .

technical hurdles. WASP carefully maps allele-specific reads, corrects for incorrect heterozygous genotypes and other sources of bias, and models overdispersion of sequencing reads. Finally, by integrating allele-specific information into a QTL mapping framework WASP attains greater power than standard QTL mapping approaches.

2.3 Unbiased read alignment with the WASP mapping tool

Mapping of reads to a reference genome is biased by sequence polymorphisms [27]. Reads which contain the non-reference allele may fail to map uniquely or map to a different (incorrect) location in the genome [27]. A common approach is to map to a ‘personalized’ genome where the reference sequence is replaced by non-reference alleles that are known to be present in the sample[30]. However, personalized genomes do not fully address the mapping problem because the genomic locations that are uniquely mappable in the reference and non-reference genome sequences differ (Figure 2.1). While these type of errors may only affect a small number of sites, they comprise a large fraction of the most significant results when tests of allelic imbalance are performed genome-wide. Genomic DNA sequencing reads can also be used to control for mapping bias, however this method reduces power to detect allelic imbalance[31].

WASP uses a simple approach to overcome mapping bias that can be readily incorporated into any read mapping pipeline. First, reads are mapped normally using a mapping tool selected by the user; mapped reads that overlap single nucleotide polymorphisms (SNPs) are then identified. For each read that overlaps a SNP, its genotype is swapped with that of the other allele and it is re-mapped. If a re-mapped read fails to map to exactly the same location, it is discarded (Figure 2.2). Unknown polymorphisms in the sample are not considered but will typically have little effect since the tests of allelic imbalance are only performed at known heterozygous sites.

2.3.1 Details on using WASP for mapping

In the WASP mapping pipeline, the user first maps reads to the genome using any mapper that outputs BAM or SAM format (Figure 2.2). For example, ChIP-seq reads can be mapped by BWA[32] or Bowtie 2[33], and RNA-seq reads can be mapped using tophat[34]. WASP then identifies mapped reads that overlap with known polymorphisms. For each read that overlaps a polymorphism, all possible allelic combinations that differ from the original read are generated and re-mapped to the genome. For example, when a read overlaps two bi-allelic SNPs, four allelic combinations are possible, three of which differ from the original read. The original read is discarded if any of the allelic combinations map non-uniquely or map to another location. Reads which overlap insertion or deletion polymorphisms are currently discarded by WASP.

This simple method has the advantages that it works with almost any existing mapping pipeline and it handles reads with sequencing errors, which are a major source of biased mapping[27, 28].

2.3.2 Comparing WASP mapping to N-masked and personal genome mapping

We performed a simulation to assess the impact of unknown polymorphisms and found that the proportion of heterozygous sites with biased mapping is very small. We simulated 100 bp reads from a lymphoblastoid cell line (NA18505) that has been genotyped by the 1000 Genomes and HapMap projects. We additionally imputed and phased genotypes for this cell line with IMPUTE2 [35] using the 1000 Genomes Phase1 integrated version 3 reference panel[12].

For each test, we evaluated the performance of WASP compared to mapping to a personal or N-masked genome. To map to personal genomes we used AlleleSeq[30]. We first created maternal and paternal reference genomes for NA18505 using the phased genotypes. We then ran the AlleleSeq pipeline using bowtie-1.1.1 [33] with –best –strata -v 2 -m 1 options as suggested by

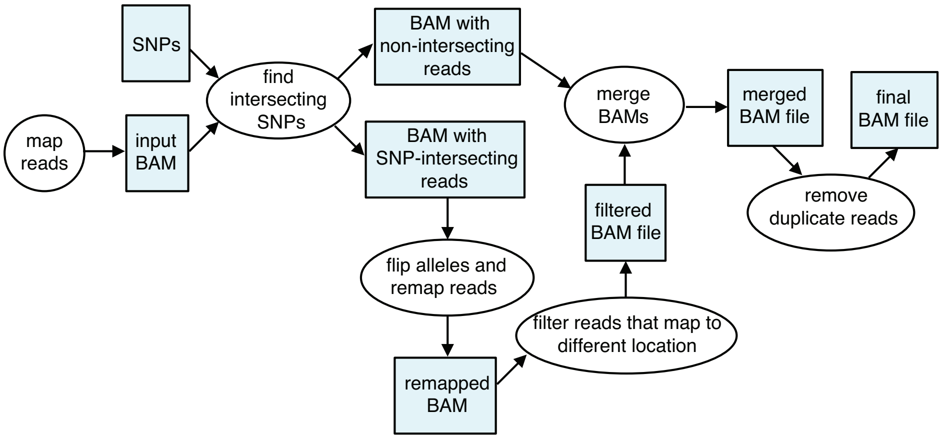

Figure 2.2: The WASP mapping pipeline. Reads are first mapped to the genome using a mapping tool of the user's choice. The aligned reads are provided to WASP in SAM (sequence alignment/map) or BAM (binary alignment/map) format, along with a list of known polymorphisms. WASP identifies reads that overlap known polymorphisms, flips the alleles in the reads, and remaps them to the genome. Reads that map to a different location than the original read are then discarded. Finally, WASP can optionally remove reads that map to the same genomic location ("duplicate reads") without introducing a reference bias.

the AlleleSeq manual. To create an N-masked genome, we created a copy of the hg19 genome with Ns in place of known variants from the NA18505 cell line. We mapped the simulated reads to the N-masked and original versions of the hg19 genome with BWA [32] allowing up to 2 mismatches per read (`-n 2`), and excluding gapped alignments (`-o 0`). The reads mapped to the original genome were provided as input to WASP. If it mapped to both genomes, we kept the location with the highest mapping quality (ties were broken randomly).

2.3.3 Quantifying the fraction of reads showing imbalance

We first identified each base where a read starting at that base would overlap a heterozygous site. We generated reads from each haplotype while introducing identical sequencing errors at a predefined rate. For each mapping type, we considered the mapping of a read to be biased if the read from one haplotype mapped to the correct location but the other did not. While reads mapped to the N-masked and personalized genomes were substantially biased and gave rise to a large number of false positives, reads mapped using WASP were almost perfectly balanced 2.3.

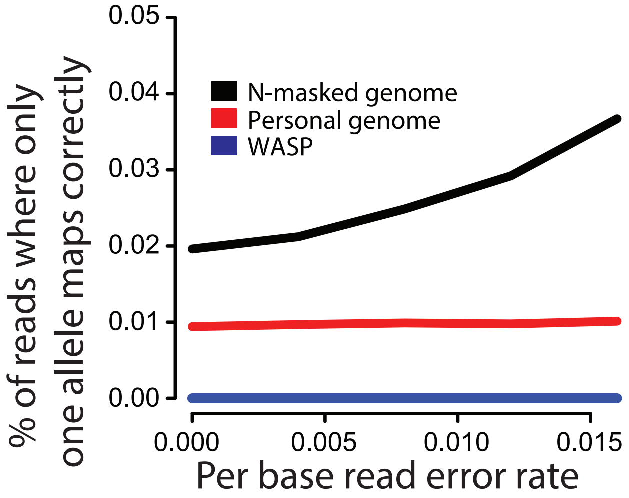

Figure 2.3: The percentage of simulated 100 bp reads at heterozygous sites where a read with one allele maps correctly and the corresponding read with the other allele does not. Reads were simulated with sequencing errors introduced at several different rates.

2.3.4 Determining the effects of unknown single nucleotide variants

One limitation of WASP is it's reliance on accurate variant information for knowing which reads to remap. With current genotyping, we are likely to miss some polymorphisms, particularly those that are only found in a small number of individuals. We tested how unknown single nucleotide variants (SNVs) affect the performance of WASP. We simulated reads from each haplotype at heterozygous sites while introducing untyped SNVs at a defined rate. We then computed the fraction of reads where the read from one allele maps correctly but the other read does not after filtering reads using WASP (Figure 2.4). The fraction of reads that map incorrectly is already very low when the rate of unknown SNVs is below 2×10^{-4}. The true rate of unknown SNVs per sample is likely to be less than 5×10^{-5} [36].

2.3.5 Assessing the effects of mapping bias on an allele-specific study

For each heterozygous site, we simulated 100 reads (of length 100 bp and with a per-base error rate of 0.01) from random bases that overlap the chosen SNP. We chose the haplotype of each simulated read at random. Reads from peaks without effects came from haplotype 1 vs haplotype

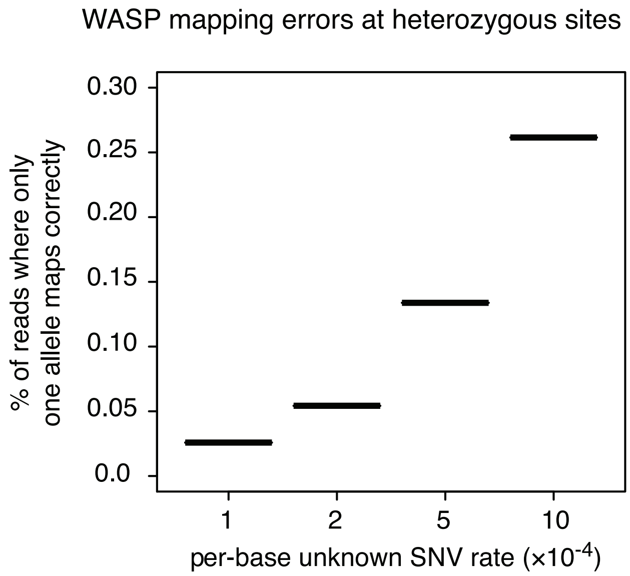

Figure 2.4: WASP mapping errors at heterozygous sites as a function of the rate of unknown single nucleotide variants (SNVs).

2 with a 1:1 ratio. Reads from peaks with effects were simulated with ratios ranging from 1.3:1 to 2.5:1 to test a range of effect sizes.

For each effect size, we simulated sets of peaks that were composed of 90% null peaks and 10% peaks with effects. We mapped the reads using each mapping scheme and performed a binomial test for imbalance on each peak, calling a locus significantly imbalanced if the p-value from the test was beneath a 10% false discovery rate (FDR) threshold. For the personal genome mapping, we used the p-values provided by the AlleleSeq pipeline. Finally, we assessed the fraction of significant loci that came from the null peaks. In the absence of imbalance caused by mapping artefacts, this should be 10%. (Figure 2.5)

2.3.6 Reads filtered by WASP

WASP filters a read when it overlap one or more SNPs and the read maps to a different genomic location (or fails to map) when the allele(s) present in the read are flipped (all possible combinations of alleles are considered). In addition, WASP currently discards all reads which overlap insertions/deletions that are polymorphic in the sample of individuals provided. We evaluated

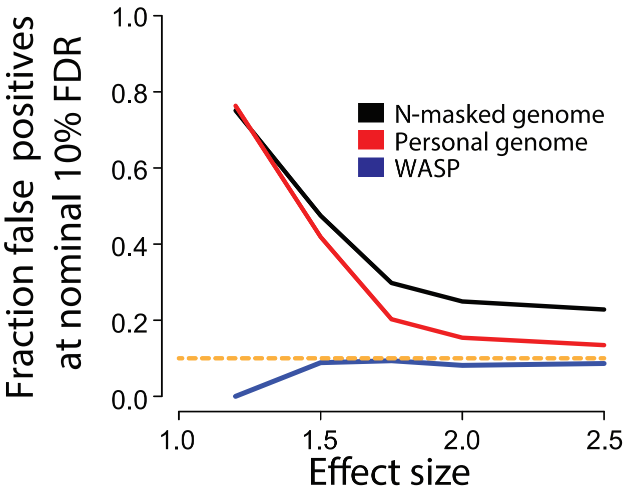

Figure 2.5: The fraction of false-positives as a function of the effect size using a nominal Benjamini-Hochberg false-discovery rate of 10%. We simulated 100 bp allele-specific reads under null (odds ratio = 1) and alternative models (odds-ratio > 1) of allelic imbalance at heterozygous sites in the genome. 90% and 10% of sites were assumed to be null and alternative sites respectively. We mapped reads using WASP, personal-genome (AlleleSeq) or N-masked-genome mapping strategies and called allele-specific sites using a binomial test.

how many reads are filtered by WASP using RNA-seq reads from a panel of 69 individuals[15] (Table 2.1). Reads were mapped as described in Section 2.5.1.

2.3.7 Limitations of WASP mapping

One disadvantage of WASP's approach is that some reads are discarded, which can cause the overall expression level of a locus to be underestimated. Several statistical methods can recover ambiguously mapped reads [37, 38], however, they are not designed for unbiased allele-specific

Table 2.1: RNA-seq reads filtered by WASP mapping in a panel of 69 individuals. The columns give the total number of mapped reads, the number of reads filtered because they overlap an indel that is present in the sample of 69 individuals, and the number of reads that are filtered because their mapping is biased. Reads are considered to have biased mapping if they overlap SNPs and map to different genomic locations when different alleles are considered.

mapped	**indel removed**	**mapping bias removed**
903346431	65900919 (7.3%)	27787224 (3.1%)

mapping and incorporating them into WASP would be technically challenging.

2.3.8 Unbiased removal of amplification effects

WASP employs a number of techniques to remove noise and biases from mapped reads. Amplification bias is a common feature of experiments that yield libraries with low complexity (e.g. ChIP-seq). Most sequencing experiments involve some amplification step where polymerase chain reaction (PCR) is used to exponentially increase the number of cDNA material for sequencing. If a small number of fragments are present before amplification, many of the resulting sequenced reads will be from the same original fragment. This can lead to increased variance and poorly calibrated results if unchecked. To control for amplification it is common to remove 'duplicate' reads that map to the same location. However, existing tools that remove duplicate reads retain the one with the highest mapping score, which will usually match the reference14. WASP provides a tool to filter duplicate reads at random, thus eliminating reference bias from this step.

2.4 Discovery of quantitative trait loci with WASP

To discover molecular quantitative trait loci (QTLs) WASP uses a statistical test, which we call the combined haplotype test (CHT). As input, the CHT takes genotype probabilities at known SNPs as well as mapped reads from sequencing-based experiments such as ChIP-seq or RNA-seq. The CHT combines two types of information: the depth of mapped reads and the allelic imbalance of mapped reads that overlap heterozygous sites.

2.4.1 Overview

The CHT models the overdispersion of read counts (both across regions and across individuals) and accounts for variability introduced by technical variation between experiments(Figure 2.6). GC content often affects read depth in a manner that is inconsistent between sequencing experiments[15, 39]. In addition, the distribution of read depths across the genome differs from experiment to experiment. For example, ChIP-seq experiments with more efficient pull-downs tend to have more reads within peaks. WASP corrects for both of these issues by fitting polynomials to the genome-wide read counts and calculating a corrected read depth for each region. Both allele-specific and total read depth counts are more dispersed than expected under models of binomial and Poisson sampling[29, 40]. To accommodate overdispersion in the data, WASP estimates separate overdispersion parameters for each individual and genomic region used in a study. Finally, to account for any remaining unknown covariates, WASP allows principal components to be included in the model fitting procedure.

Following correction for biases described above, WASP uses a statistical test, the combined haplotype test (CHT), to identify cis-acting QTLs. The CHT tests whether the genotype of a test SNP is associated with total read depth and allelic imbalance in a target region. The CHT jointly models two components: the allelic imbalance at phased heterozygous SNPs and the total read depth in the target region. The two components of the test are linked together by shared parameters that define their effect sizes. For a target region and test SNP pair, the CHT models the expected number of reads for an individual as a function of the individual's genotype, the effect size, the GC content, additional covariates (such as principal component loadings), and the total number of mapped reads in the region (across all individuals). The probability of the observed number of reads in the target region is calculated using the expected number of reads and two overdispersion parameters.

Allelic imbalance of reads overlapping heterozygous SNPs within a target region is modeled as a function of the shared effect size parameters. The probability of the observed allele-specific read counts is then defined by the effect size and a single overdispersion parameter. We also allow for the possibility of genotyping errors by assuming that allele-specific read counts are drawn from a mixture, with a small probability that a given individual is a mistyped homozygote. WASP combines information across multiple heterozygous sites and the current implementation assumes that haplotype phasing is correct. Incorrect phasing will decrease WASP's power to detect associations but will not increase false positives.

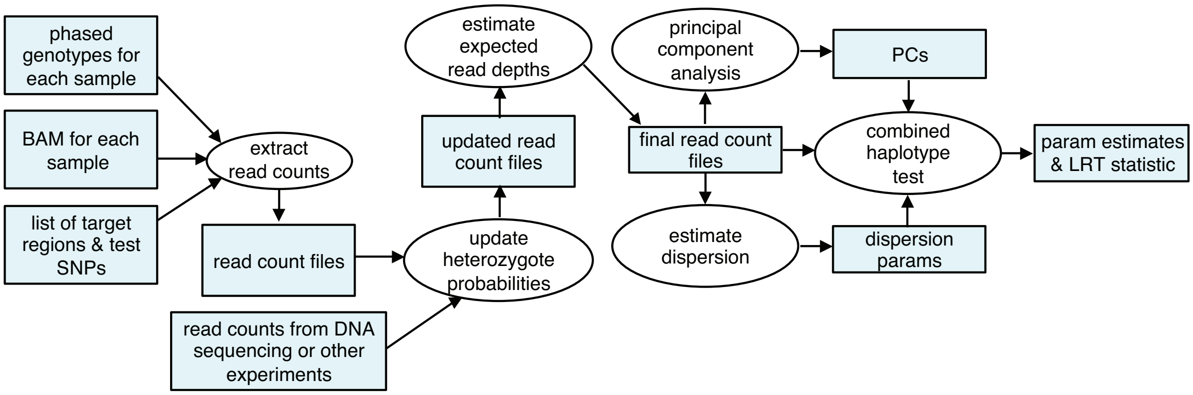
phased genotypes for each sample
BAM for each sample
list of target regions & test SNPs
extract read counts
read count files
read counts from DNA sequencing or other experiments
update heterozygote probabilities
updated read count files
estimate expected read depths
final read count files
principal component analysis
estimate dispersion
PCs
dispersion params
combined haplotype test
param estimates & LRT statistic

Figure 2.6: The WASP combined haplotype test pipeline. Mapped reads (in BAM or SAM format) for each individual, genotypes for known SNPs, and a list of regions and SNPs to test are provided to WASP. WASP extracts read counts for the target regions as well as allele-specific read counts. Read counts from multiple sources can be used to update heterozygous probabilities. Expected read counts for each region are adjusted by modeling the relationships between read counts and GC content and read counts and total read counts for each sample. Dispersion parameters are estimated from the data and provided to the combined haplotype test along with the read counts. Principal components can optionally be used as covariates by the test.

Table 2.2: Description of mathematical variables used in the combined haplotype test

Variable	Description
Index variables	
h	test number (one per test SNP / target region pair)
i	individual
j	target region
k	SNP within target region
m	test SNP
Latent variables	
α_h	molecular phenotype level of the reference allele for test h
β_h	molecular phenotype level of the alternative allele for test h
p_h	fraction of allele-specific reads expected from reference allele $(p_h = \frac{\alpha_h}{\alpha_h+\beta_h})$
$T^*_{i,j}$	genotype-independent expected total read count for individual i, target region j
λ_{hi}	expected total read count for test h, individual i
Ω_i	overdispersion of read counts for individual i (across all target regions)
ϕ_j	overdispersion of read counts for target region j (across all individuals)
Υ_i	overdispersion of allele-specific reads for individual i
Observed variables	
x_{ij}	number of reads for individual i, target region j
G_{im}	genotype call for individual i, test SNP m
T_i	total number of genome-wide mapped reads for individual i
n_{ik}	total number of allele-specific reads for individual i, target SNP k
y_{ik}	number of allele-specific reads from reference haplotype for individual i, target SNP k
H_{ik}	probability individual i is heterozygous for target SNP k

2.4.2 *The combined haplotype test details*

The combined haplotype test (CHT) determines whether the genotype of a test SNP, m, is associated with read depth and allelic imbalance within a nearby target region, j, on the same chromosome (Figure 2.6, Table 2.2). Each test is performed on a test SNP and target region pair, $h = \{m, j\}$. A target region may be discontiguous and span multiple genomic loci. For example, the exons of a gene can be used as a target region when searching for expression QTLs using RNA-seq reads. The test SNP is not required to be within the target region, but is assumed to be nearby and cis-acting. This allows us to combine information from across phased heterozygous SNPs and assign reads to one haplotype or the other.

2.4.3 The basic model

The CHT is a likelihood ratio test with two components. One component models the depth of mapped reads within the target region, and the other component models the allelic-imbalance of reads that overlap heterozygous SNPs. Both components of the test are parameterized by α_h and β_h, which define the expected read depth from chromosomes with the reference and alternative alleles. Since variants are assumed to be additive and cis-acting, the expected allelic imbalance in heterozygotes is $p_h = \frac{\alpha_h}{\alpha_h+\beta_h}$[26].

2.4.4 Modeling the read depths

The number of reads mapping to a target region is often modeled using a poisson distribution[41]. However, the poisson assumption that the variance is equal to the mean is often violated because read counts from target regions are overdispersed. Part of this overdispersion can be accommodated by modeling the data with a negative-binomial distribution with a variance parameter for each test[29]. However, the negative binomial distribution assumes that the mean and variance have a quadratic relationship that is consistent across individuals. We have found that this assumption is violated by sequencing data and causes poor calibration of the tests, particularly when sample sizes are small. The CHT therefore includes negative binomial overdispersion parameters for each individual, Ω_i, and for each target region, ϕ_j. After adding these additional dispersion parameters, the data are modeled with a beta-negative-binomial (BNB) distribution. The expected number of read counts for an individual, λ_{hi}, is defined as:

$$\lambda_{hi} = \begin{cases} 2\alpha_h T_i & \text{if } G_{im} = 0 \text{ (homozygous allele 1)} \\ \\ (\alpha_h + \beta_h)\, T_i & \text{if } G_{im} = 1 \text{ (heterozygous)} \\ \\ 2\beta_h T_i & \text{if } G_{im} = 2 \text{ (homozygous allele 2)} \end{cases} \tag{2.1}$$

where G_{im} is the genotype of individual i at test SNP m, and T_i is the total number of reads mapped genome-wide for individual i.

The likelihood of the parameters is then given by the equation:

$$\mathrm{L}\left(\alpha_h, \beta_h, \Omega_\bullet, \phi_j \left| D_h\right.\right) = \prod_i \Pr_{\text{BNB}}\left(X = x_{ij} \left| \lambda_{hi}, \Omega_i, \phi_j\right.\right) \tag{2.2}$$

where x_{ij} is the number of reads for individual i in target region j.

2.4.5 Correcting for GC content and other effects on expected read depth

Since the number of mapped reads can differ between sequencing lanes and runs, we initially model the expected number of counts, λ_{hi}, as a linear function of the total number of mapped reads for each individual, T_i. However, technical variation between experiments can change this relationship and reduce power to detect true differences in read depths between samples or cause spurious associations. As described below, we directly model some known sources of technical variation and estimate adjusted total read depths, T^*_{ij}, for each individual and target region. We then replace T_i in Equation 2.1 with T^*_{ij}. This gives us a more accurate estimate of the expected number of reads and improves our ability to detect true QTLs.

2.4.6 Adjusting total read depth

In RNA-seq experiments, a large fraction of mapped reads can come from a small number of highly expressed genes. Variation in the expression level of these genes can therefore have a large effect on the number of reads that map to all other genes[42]. In ChIP-seq experiments, the fraction of reads that come from peaks varies between experiments, likely due to differences in the efficiency of immuno-precipitation (Figure 2.7).

To account for these types of variation, we calculate an adjusted total read depth, T^*_{ij} for each region and individual. The adjusted read depth is defined by a quartic function of the total read depth (summed across individuals) for each target region. We estimate the coefficients of the quartic function separately for each individual using a maximum likelihood approach described below (Figure 2.7).

2.4.7 GC content correction

GC content also affects read depth, with a relationship that varies across samples [15, 39]. For example, in some samples, high GC content regions have high read depth, while in other samples they have low read depth. To account for this variation, we add GC content terms to the model of adjusted total read depth. These terms are modeled with a log linker so that T^*_{ij} is guaranteed to be positive. After fitting this model we can calculate an adjusted total read depth for each region that takes into account both the GC content variation and the total read depth variation (Figure 2.7).

2.4.8 Fitting adjustment coefficients

For each target region, j, we count the total number of reads $v_j = \sum_i x_{ij}$ and calculate the GC content w_j. Then, for each individual i, we find maximum likelihood estimates of coefficients $a_{0i}, a_{1i}, \ldots, b_{4i}$ that define the adjusted expected counts, T^*_{ij}:

$$\mathrm{L}\left(a_{0i}, a_{1i}, \ldots, b_{4i} \left| D_i \right.\right) = \prod_j \Pr_{\text{Pois}} \left(X_{ij} = x_{ij} \middle| T^*_{ij} \right) \tag{2.3}$$

$$T^*_{ij} = \exp\left(a_{0i} + a_{1i} w_j + a_{2i} w_j^2 + a_{3i} w_j^3 + a_{4i} w_j^4\right) \left(b_{1i} v_j + b_{2i} v_j^2 + b_{3i} v_j^3 + b_{4i} v_j^4\right) \tag{2.4}$$

2.4.9 Modeling the allelic imbalances

Allele-specific read counts are sometimes modeled using the binomial distribution [43], however, we have found that allele-specific read counts are overdispersed. We instead model allele-specific read counts with a beta-binomial (BB) distribution and include a parameter Υ_i (estimated separately) that captures the overdispersion for each individual. The likelihood of the parameters given the data is then:

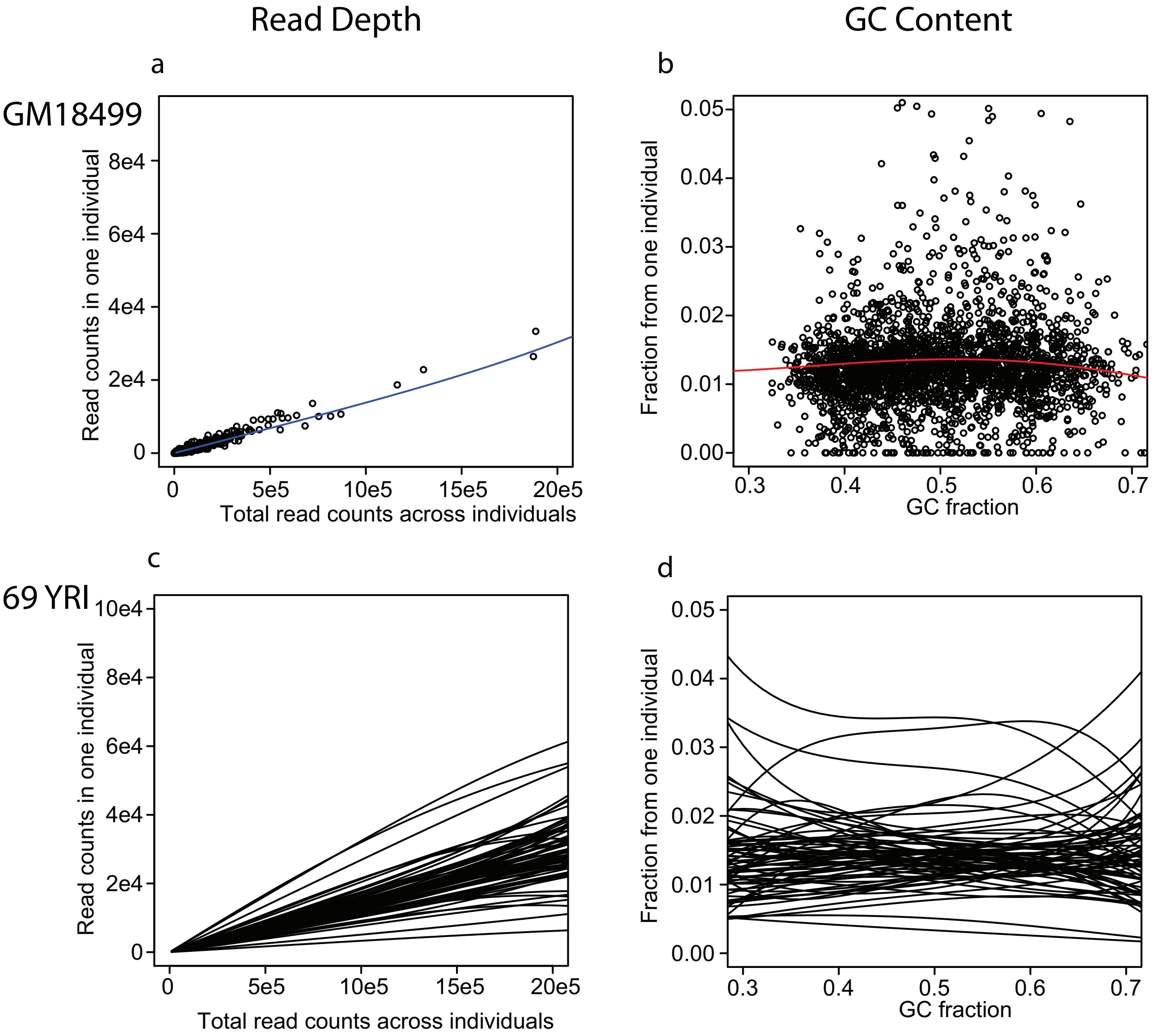

Figure 2.7: Adjusting expected read counts based on total read depths and GC content. (**a**) H3K27ac ChIP-seq read counts in target regions from cell line GM18499 as a function of the total number of reads across all individuals in the same target regions. The blue line shows the fitted quartic function used to adjust expected read depths. (**b**) H3K27ac read counts in target regions from cell line GM18499 as a function of GC content. The red line shows a the fitted quartic function used to adjust expected read depths. (**c**) Fitted functions for all 69 Yoruba individuals showing the relationship between total and per-individual read counts. (**d**) Fitted functions for all 69 Yoruba individuals showing the relationship between read counts and GC content.

$$\mathrm{L}\left(\alpha_h, \beta_h \left| D\right.\right) = \prod_i \prod_k \Pr_{\mathrm{BB}}\left(Y = y_{ik} \left| n_{ik}, p_h, \Upsilon_i\right.\right) \tag{2.5}$$

where y_{ik} is the number of allele-specific reads from the reference haplotype and n_{ik} is the total number of allele-specific reads for individual i at target SNP k. The expected fraction of allele-specific reads from the reference allele is $p_h = \frac{\alpha_h}{\alpha_h + \beta_h}$.

2.4.10 Correcting for incorrect genotype calls

SNP genotypes that are incorrectly called as heterozygous are a major source of false positives, since reads that overlap them appear to come from only one allele. To account for this issue, we assume that allele-specific reads are drawn from a mixture of two beta-binomials, with probabilities H_{ik} and $1 - H_{ik}$, where H_{ik} is the probability that individual i is heterozygous for SNP k. Reads from heterozygous individuals contain the reference allele with probability p_h. We assume that reads from homozygous individuals still have a small probability of coming from the other allele due to sequencing errors, which occur with probability, p_{err}. The probability of observing y_{ik} reads from the reference allele for individual i at SNP k then becomes:

$$\begin{aligned} \Pr_{\mathrm{BB-mix}}\left(Y = y_{ik} \left| p_h, n_{ik}, \Upsilon_i, H_{ik}\right.\right) = H_{ik} \Pr_{\mathrm{BB}}\left(Y = y_{ik} \left| p_h, n_{ik}, \Upsilon_i\right.\right) \\ +(1 - H_{ik}) \left[\Pr_{\mathrm{BB}}\left(Y = y_{ik} \left| p_{\text{err}}, n_{ik}, \Upsilon_i\right.\right) + \Pr_{\mathrm{BB}}\left(Y = y_{ik} \left| 1 - p_{\text{err}}, n_{ik}, \Upsilon_i\right.\right)\right] \end{aligned} \tag{2.6}$$

We found that even SNPs with heterozygous probabilities of 1.0 are occasionally miscalled so we set heterozygous probabilities to a maximum value of 0.99. We then update this heterozygous probability using sequencing data obtained from the same individual. Sequencing data may consist of DNA sequencing reads or reads aggregated across multiple types of experiments performed on the same individual (e.g. RNA-seq and ChIP-seq reads).

For a SNP with heterozygous probability $H_{ik} = \min(0.99, H_{ik}^{\text{obs}})$, we define the updated heterozygous probability, $\hat{H}_{ik}$ as:

$$\hat{H}_{ik} = \frac{H_{ik}\,\mathrm{Pr}_{\mathrm{Bin}}\left(D\,|p=0.5\right)}{H_{ik}\,\mathrm{Pr}_{\mathrm{Bin}}\left(D\,|p=0.5\right) + \left(1-H_{ik}\right)\left[\mathrm{Pr}_{\mathrm{Bin}}\left(D\,|p=p_{err}\right) + \mathrm{Pr}_{\mathrm{Bin}}\left(D\,|p=1-p_{\mathrm{err}}\right)\right]} \tag{2.7}$$

2.4.11 The combined likelihood ratio test

The combined likelihood of both components of the model is:

$$\mathrm{L}\left(\alpha_h, \beta_h, \phi_j\,|D\right) = \prod_i \left[\Pr_{\mathrm{BNB}}\left(X = x_{ij}\,|\lambda_{hi}, \Omega_i, \phi_j\right) \prod_k \Pr_{\mathrm{BB-mix}}\left(Y = y_{ik}\,\middle|\,p_h, n_{ik}, \Upsilon_i, \hat{H}_{ik}\right)\right] \tag{2.8}$$

To test for an association with genotype we perform a likelihood ratio test that compares the alternative hypothesis $\alpha_h \neq \beta_h$ to the null hypothesis $\alpha_h = \beta_h$. The CHT returns a likelihood ratio statistic $\Lambda = \frac{\mathrm{L}(\hat{\theta}_1|D)}{\mathrm{L}(\hat{\theta}_0|D)}$ where $\hat{\theta}_1$ and $\hat{\theta}_0$ are maximum likelihood estimates of the parameters under the alternative and null hypotheses. P-values can be calculated from the the test statistic under the asymptotic assumption that $-2\log(\Lambda)$ is χ^2 distributed with one degree of freedom.

2.4.12 Estimating overdispersion parameters

In order to estimate the genome-wide overdispersion parameters Ω_i and Υ_i, we use the same likelihood equations as in the CHT, but assume that there are no genetic effects. This means that for the read depth part of the test λ_{hi} is equal to the expected counts T^*_{ij}, and for the allele-specific part of the test p_h is equal to 0.5. Since the allele-specific and read depth parts of the likelihood equation are independent, we can fit the overdispersion parameters separately.

2.4.13 Beta-Negative-Binomial parameter estimation

To find the maximum likelihood estimate of Ω_i we need to sum the log likelihood across all regions. This presents a problem, as ϕ_j must also be estimated for each region. We therefore iteratively estimate ϕ_j by first finding a maximum likelihood estimate for ϕ_j for each region using the equation:

$$\mathrm{L}\left(\phi_j \left| D\right.\right) = \prod_i \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \left| \lambda = T^*_{ij}, \Omega_i, \phi_j \right. \right) \right] \tag{2.9}$$

and then finding a maximum likelihood estimate for Ω_i for each individual using the equation:

$$\mathrm{L}\left(\Omega_i \left| D\right.\right) = \prod_j \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \left| \lambda = T^*_{ij}, \Omega_i, \phi_j \right. \right) \right] \tag{2.10}$$

We repeat this iterative procedure until the improvement in the likelihoods becomes negligible.

2.4.14 Beta-Binomial parameter estimation

We calculate the genome-wide likelihood of Υ_i by taking the product of likelihoods from all target region SNPs that are heterozygous in individual i. We again assume there is no genetic effect, so p = 0.5, and we use the following equation to find the maximum likelihood estimate of Υ_i:

$$\mathrm{L}\left(\Upsilon_i \left| D\right.\right) = \prod_k \Pr_{\mathrm{BB-mix}} \left(Y = y_{ik} \left| n_{ik}, p = 0.5, \Upsilon_i, \hat{H}_{ik} \right. \right) \tag{2.11}$$

2.4.15 CHT calibration

Generally the overdispersion parameters estimated by the CHT allow the model to be well calibrated, showing little signal when run on permuted data. However permuted tests can sometimes diverge from the null, particularly when small sample sizes are used. This may occur because by chance the permutations are unable to completely break up the signal when there aren't many samples to permute or because of inaccuracy in the overdispersion estimates. We suggest running the CHT on permuted data using the options we provide and visualizing the results with a quantile-quantile plot to ensure that the test is working properly. If the permutations do not follow the null, the user may manually set overdispersion parameters or adjust the p-values according to

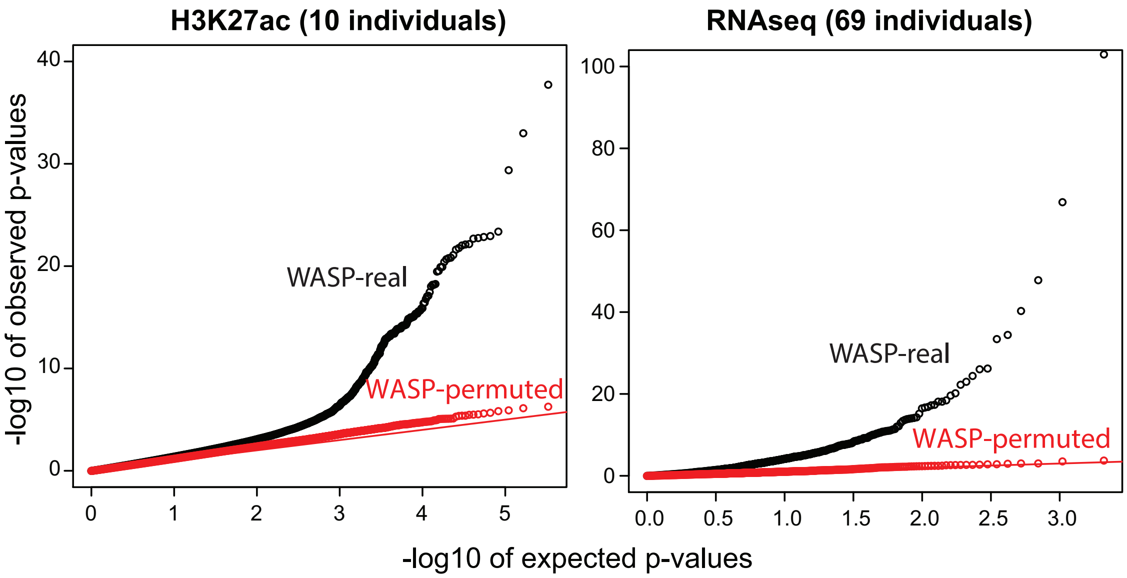

Figure 2.8: Quantile-quantile plots of ranked -log10 p-values from the combined haplotype test. The permuted points are for same datasets but with the genotypes of each SNP shuffled. (**a**) Ranked -log10 p-values from running the combined haplotype test on H3K27ac ChIP-seq data from 10 lymphoblastoid cell lines compared to p-values expected under the null hypothesis. (**b**) Ranked -log10 p-values from running the combined haplotype test on RNA-seq data from 69 YRI cell lines. The test was run only on eQTLs that were previously identified in cell lines derived from European individuals[44].

the permuted distribution.

2.4.16 Correcting for unknown covariates using principal components

Both known and unknown covariates such as time of experiment, age of sample, etc. can affect molecular trait measurements and confound QTL studies. Principal component analysis (PCA) is sometimes used to capture and remove these effects [15, 22]. To leverage PCA while maintaining the discrete nature of the count data, the CHT directly models the covariate effects. To do this we include a user-defined number of PCA loadings $u_{i\bullet}$ and fit coefficients $c_{h\bullet}$ when calculating λ_{hi}.

$$\lambda_{hi} = \begin{cases} 2\alpha_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 0 \text{ (homozygous allele 1)} \\ \\ (\alpha_h + \beta_h)\,(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 1 \text{ (heterozygous)} \\ \\ 2\beta_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 2 \text{ (homozygous allele 2)} \end{cases} \tag{2.12}$$

Fitting many coefficients simultaneously can be quite slow, but since the principal components are by definition orthogonal, we can optimize their coefficients one at a time without losing accuracy. We then use the fitted coefficients to calculate λ_{hi} for the null and alternative models.

2.5 WASP combined haplotype test performance evaluation

To evaluate the performance of WASP, we tested the ability of the combined haplotype test at (i) recalling in 69 individuals QTLs that were previously identified in a different population, (ii) calling novel QTLs genome-wide using data from H3K27ac ChIP-seq experiments that were performed in 10 LCLs [40], and (iii) calling QTLs from simulated data against other alllele specific QTL calling software.

2.5.1 Identifying known European eQTLs in 69 Yoruba LCLs

We downloaded eQTLs which were identified in 373 European lymphoblastoid cell lines (LCLs) by the GEUVADIS project [44]. We identified a subset of 2098 of these eQTL SNPs that were segregating in an independent dataset of 69 Yoruba LCLs [15] with a minimum minor allele count of 2. We mapped RNA-seq reads from the 69 Yoruba LCLs to the hg19 genome using tophat with the options `--segment-length 17`, `--b2-sensitive` and `--no-coverage-search` and processed the mapped reads with the WASP mapping pipeline. We applied the CHT and linear model to the mapped RNA-seq reads. WASP discovers 627 of the eQTLs at a false discovery rate (FDR) of 10%, which is impressive considering (1) our smaller sample size, (2) that some fraction of the original eQTLs are false positives, and (3) that some of the European eQTLs will be absent or at very low frequency in the Yoruba. This number increases to 673 when 5 principal components are included as covariates. By comparison, when we adopt a standard eQTL discovery method (linear regression on quantile normalized and GC-corrected data), we identify only 446 eQTLs (617 when 5 principal components are included as co-variates). P values obtained by running the CHT on the same dataset with permuted genotypes do not depart substantially from the null expectation, indicating that the test is well-calibrated. (Figure 2.9). We also examined the correlation between the allelic imbalance estimate from CHT and the reported genotype-expression correlation from GEUVADIS (Figure 2.10). The correlation is strongest at eQTLs that are close to the transcription start site (Spearman's $\rho = 0.72$, $p = 7 \times 10^{-56}$) and decreases within increasing distance (Figure 2.10). This is likely because the current implementation of WASP assumes that haplotype phasing is correct but phasing accuracy decreases with distance.

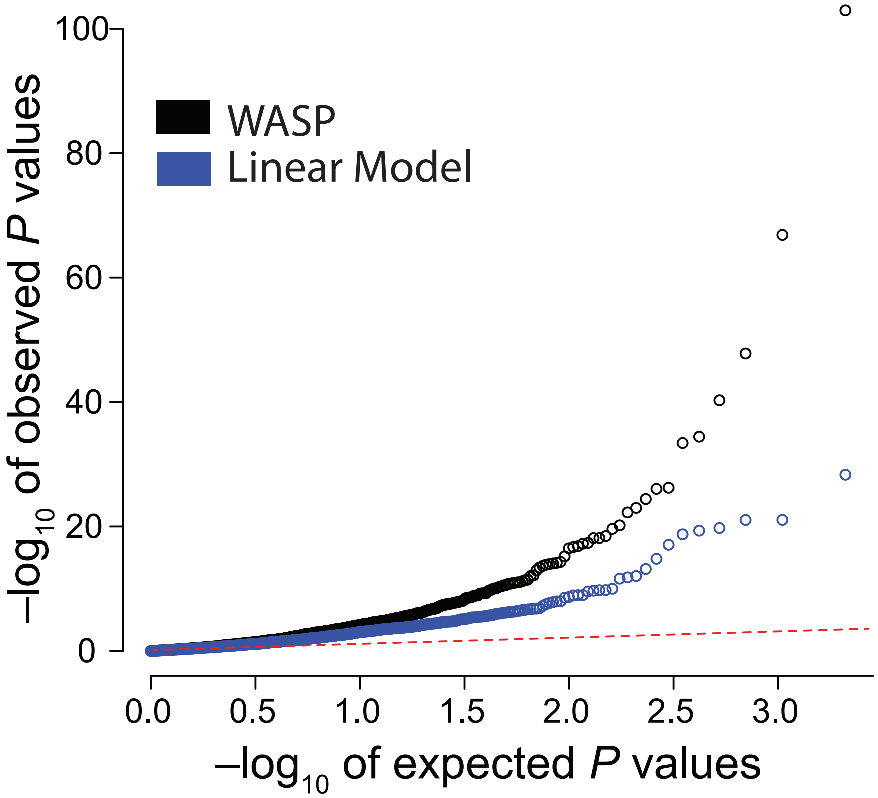

Figure 2.9: Identifying European eQTLs from the GEUVADIS consortium using an independent dataset of RNA-seq from 69 Yoruba lymphoblastoid cell lines.

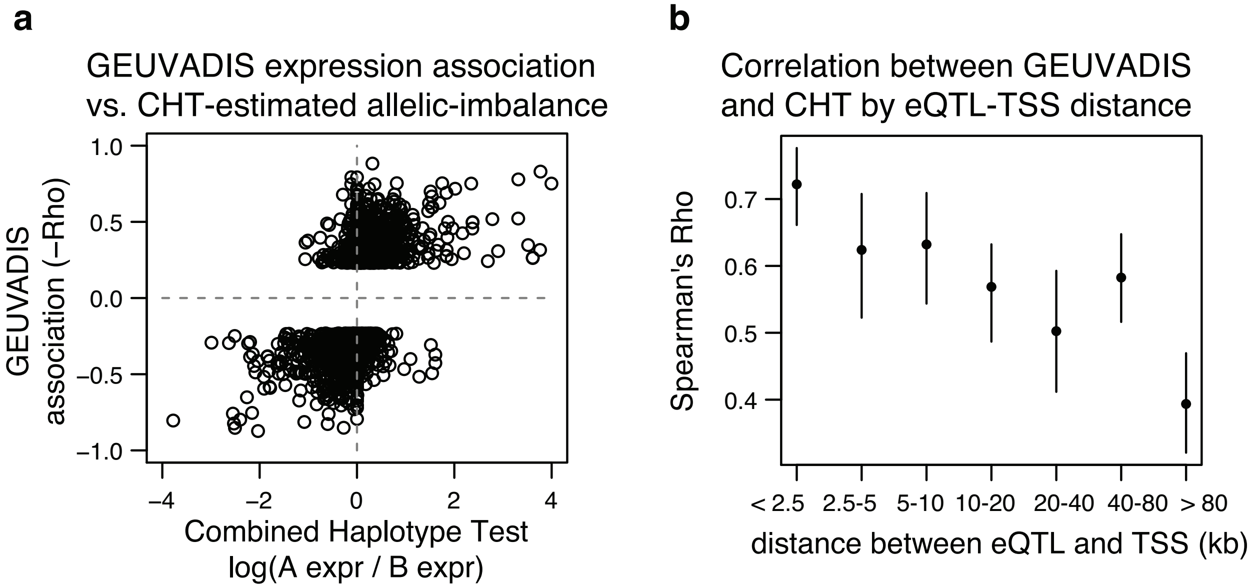

Figure 2.10: Comparison of results from GEUVADIS to allelic imbalance estimates from the Combined Haplotype Test (CHT). We ran CHT on RNA-seq data from 69 Yoruba cell lines and compared the estimated allelic imbalance to the genotype-expression associations reported by GEUVADIS. The comparison was performed at GEUVADIS eQTLs that were identified in European cell lines [44]. (**a**) Scatter plot showing the GEUVADIS-reported association statistic (Spearman's ρ) versus the allelic imbalance estimate from CHT. (**b**) Correlation between GEUVADIS-reported association and CHT's estimate of allelic imbalance as a function of distance between the eQTL and the transcription start site (TSS) of the associated gene. Whiskers are 95% confidence intervals from 1000 bootstraps.

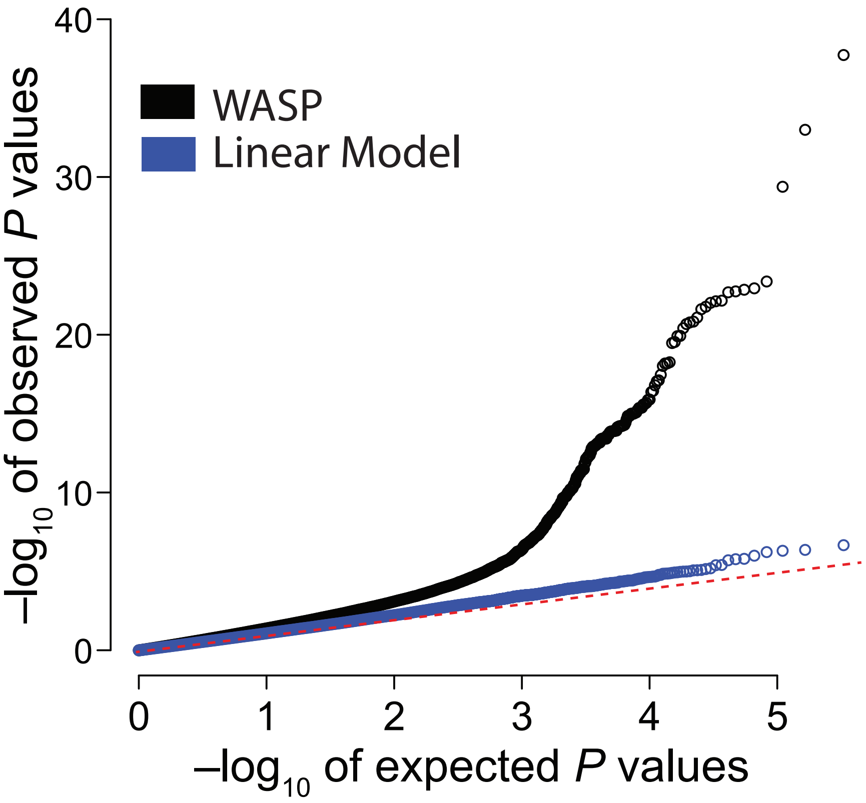

Figure 2.11: Identification of novel QTLs using H3K27ac ChIP-seq data from 10 Yoruba lymphoblastoid cell lines.

2.5.2 Genome-wide QTL discovery in small sample sizes of ChIP-seq data

We also applied the two models to a dataset of ChIP-seq data for the histone modification H3K27ac from 10 individuals, which we collected in a previous study [40]. We mapped the ChIP-seq reads to the hg19 genome using the default options of bowtie2 and processed the mapped reads with the WASP mapping pipeline. Principal components were not included in this analysis because of the small number of dimensions in the dataset. As test SNPs we chose SNPs that were segregating in the 10 individuals and defined the target region as a 2 kb region centered on the test SNP. We only tested target regions with at least 100 filtered reads summed across individuals (Figure 2.11).

2.5.3 Comparing CHT to other QTL mapping strategies using simulations

Read count over-dispersion and genotyping errors can lead to artifacts when testing for QTLs. Tests that do not account for these problems may appear to identify more QTLs simply because they identify more false positives. Since is difficult to distinguish between true effects and artifacts in real data, we used simulations to compare the relative sensitivity of the CHT and several other methods for QTL discovery.

2.5.4 Simulating read depth and allele-specific counts

We simulated genotypes for individuals with a minor allele frequency of 0.2 and discarded simulated sites with fewer than 2 heterozygous individuals. We then simulated total read counts by observing a beta negative binomial random variable with the following dispersion parameters: $\Omega = 0.01$ and $\phi_j = 100$. These parameter values were chosen to be similar to our dispersion estimates from real data

The mean for the distribution, λ, was based on the simulated genotype, G, the effect size, E, and whether the minor allele has higher ($\delta = 1$) or lower mean count ($\delta = 0$). In our simulation we randomly set δ to 0 or 1 with equal probability.

$$\lambda = \begin{cases} 200 & \text{if } G = 0 \text{ (homozygous major)} \\ \\ 200\,(2+E)\,\delta + 200\left(\frac{2}{2+E}\right)(1-\delta) & \text{if } G = 1 \text{ (heterozygous)} \\ \\ 200\,(2+2E)\,\delta + 200\left(\frac{2}{2+2E}\right)(1-\delta) & \text{if } G = 2 \text{ (homozygous minor)} \end{cases} \tag{2.13}$$

For heterozygous individuals, we simulated allele-specific read counts by drawing from a beta binomial distribution with the following parameters: $n = 20$, $p = \frac{1}{1+E}\delta + \frac{E}{1+E}(1-\delta)$, and $\Upsilon = 0.2$. To simulate errors in genotyping, 1% of the counts were drawn from a beta binomial

Table 2.3: Summary of QTL methods tested

Method	Description
CHT	Our method. Combines allele-specific (beta binomial) and read depth (beta negative binomial) information.
TReCASE	Combines allele-specific (beta binomial) and read depth (negative binomial) information [26].
Regression	Simple linear regression
Beta Binomial	A likelihood ratio test for imbalance in allele-specific read counts similar to that described in [45]
Kruskal-Wallis	Non-parametric test for association using read depth only.

distribution with $p = 0.99$, representing a target SNP that was labeled as heterozygous but was actually homozygous.

2.5.5 Comparing QTL model sensitivities

We compared five methods for QTL discovery, which are summarized in Table 2.3.

We simulated 10,000 sites under the null ($E = 0$) and alternative hypotheses (E varied). We then compared the performance of the tests summarized in Table 2.3 using receiver operating characteristic (ROC) curves (Figure 2.12). For the smaller sample sizes (10 or 20 individuals), CHT outperforms all other tests. Interestingly for sample size 10, simple regression outperforms TReCASE likely because linear regression can more flexibly model the variance, which helps it avoid false positives. For larger sample sizes, CHT and TReCASE perform similarly and both outperform regression. The beta binomial and Kruskal-Wallis tests perform relatively poorly under all conditions. Like the CHT, TreCASE uses both allelic imbalance and read depth information, however it does not account for overdispersion, genotyping errors, or biased mapping, which increase the false positive rate when using real data.

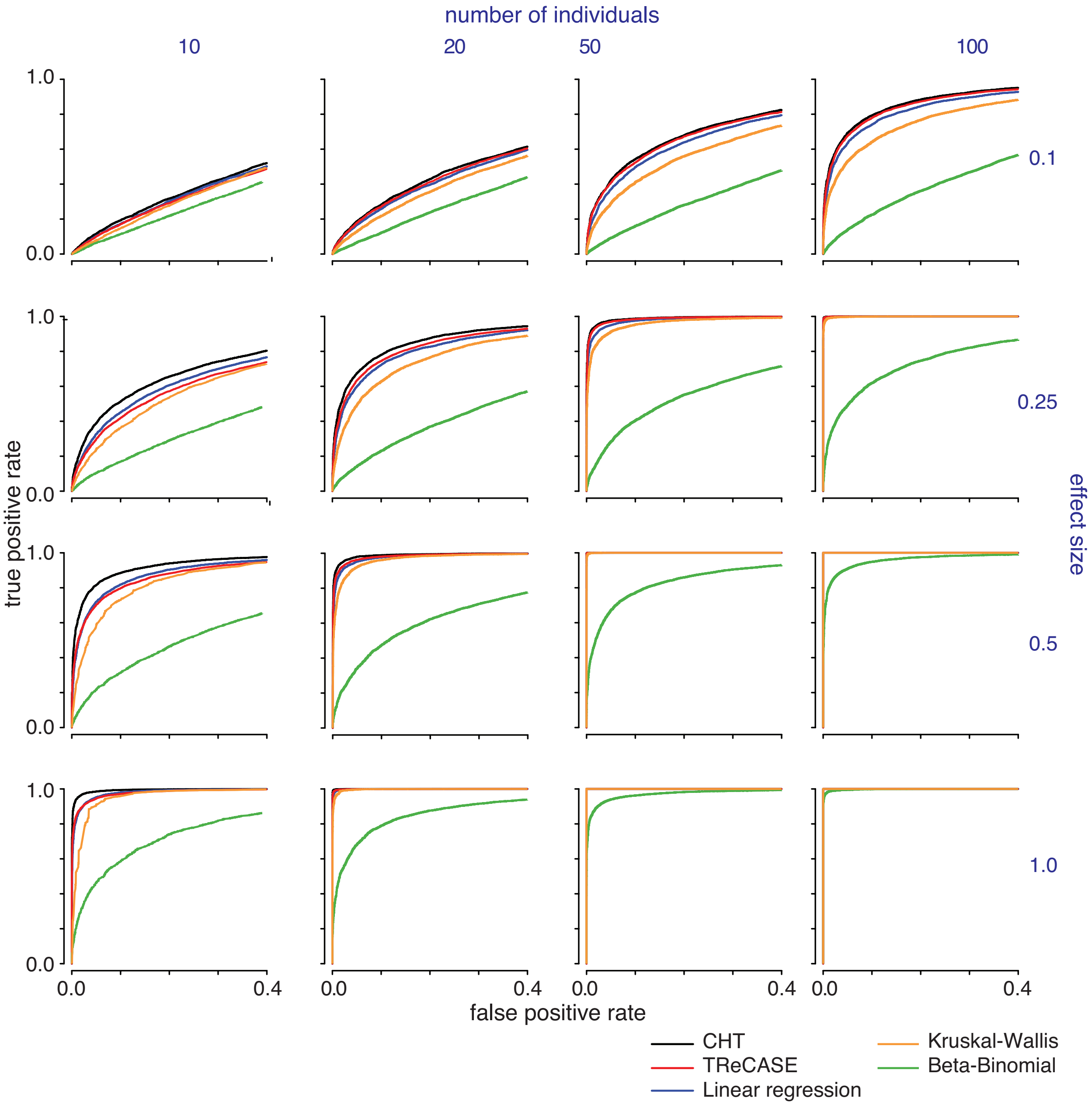

Figure 2.12: Receiver operating characteristic curves (ROC) showing the performance of five methods for QTL identification on simulated data. The panels show performance for different numbers of individuals and effect sizes.

2.6 Testing effects of reduced allelic imbalance

The CHT combines allele-specific and read depth information by assuming $p = \frac{\alpha}{\alpha+\beta}$. Previous work suggests that this assumption is reasonable for most eQTLs[15], however under some circumstances QTLs may have buffered or non-additive effects. To test how non-additive or buffered genotypic effects change the CHT's power to detect QTLs, we simulated read count data under a model of allele-specific buffering.

2.6.1 Simulating sites with reduced allelic imbalance

We simulated read depth and allele-specific data using the methods described in Section 2.5.4, but with the addition of an allele-specific buffering parameter, κ. We then redefined the allelic imbalance parameter as $p = \frac{1}{1+E_{AS}}$, where $E_{AS} = \kappa E$.

2.6.2 Results

We again performed simulations as described in Section 2.5.4, but introduced the allele-specific buffering parameter, κ, when simulating read counts under the alternative hypothesis. We simulated reads using the following values of κ: 1.0 (no buffering), 0.75, 0.50, and 0.25. As expected, the performance of the CHT is worse for lower values of κ because allelic imbalance is attenuated. Under most conditions the CHT still outperforms a simple regression if κ is greater than 0.5. With $\kappa = 0.25$, however, there is a modest drop-off in power (Figure 2.13).

2.7 CHT running time

To assess the computational running time of the CHT we simulated data for between 10 and 1000 individuals. To simulate data, we made copies of the H3K27ac ChIP-seq data from 10 individuals. We then obtained the mean running time per test by running the CHT on several hundred sites. The mean running time increases linearly with the number of individuals, and we found the mean

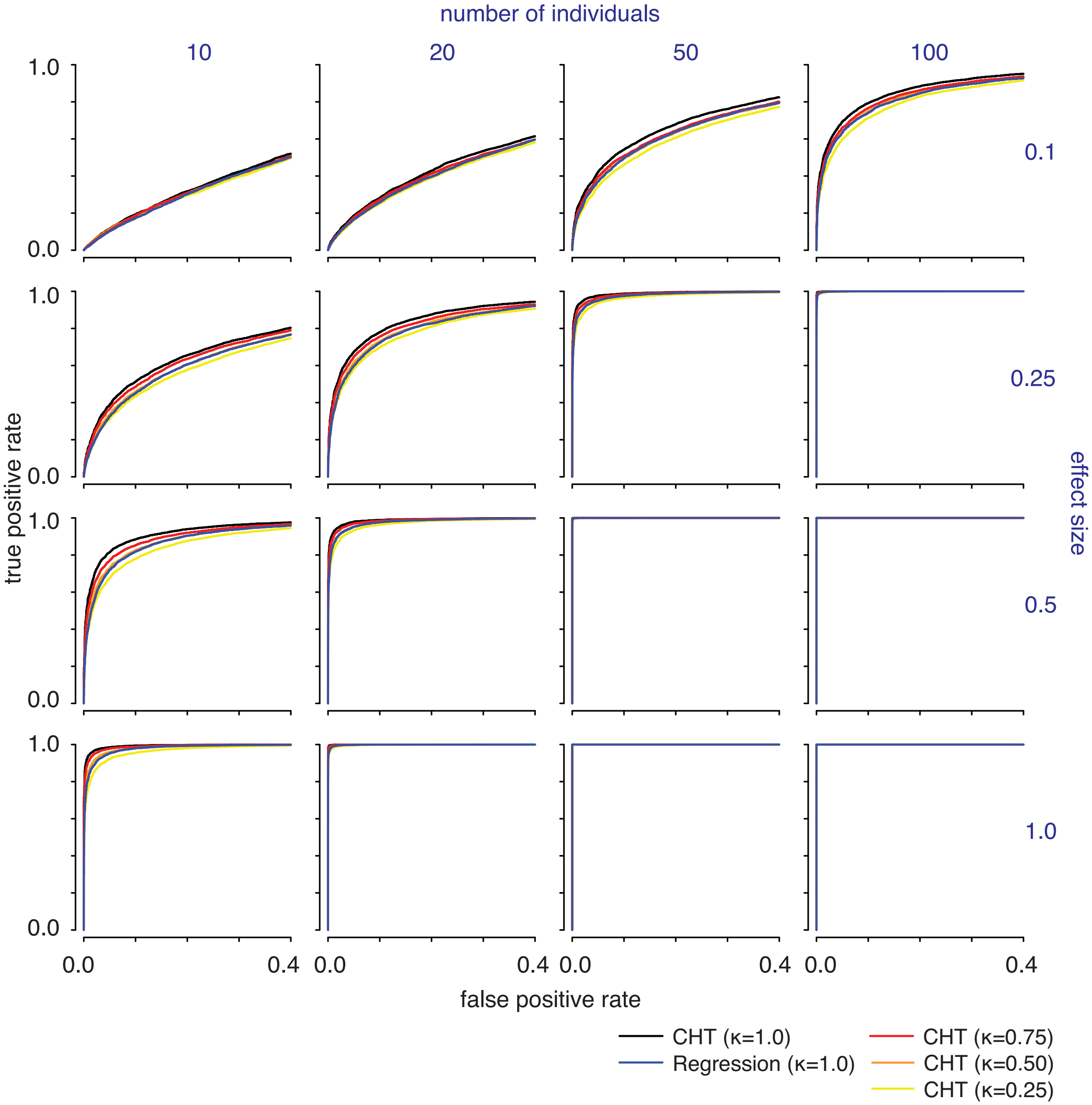

Figure 2.13: Receiver operating characteristic curves (ROC) showing the performance of CHT with different levels of allele-specific buffering. Each panel shows performance with different numbers of individuals and effect sizes. The different line colors indicate the value of the allele-specific buffering parameter κ that was used for simulating read counts under the alternative model. When $\kappa \neq 1.0$ the genotypes have non-additive effects. Results for a simple linear regression for $\kappa = 1.0$ are shown for comparison.

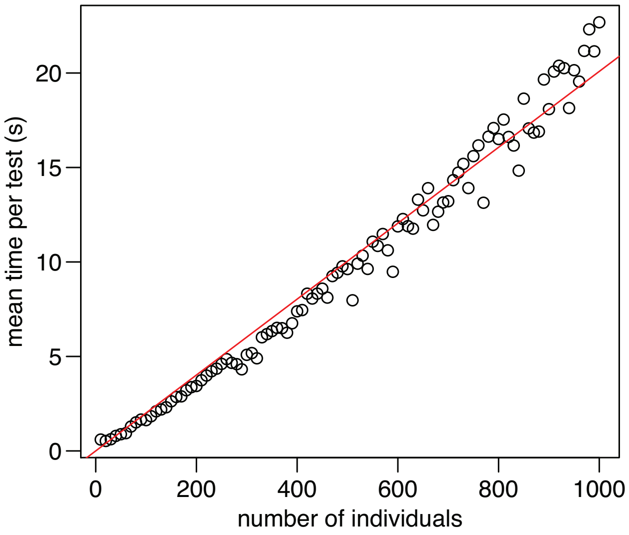

Figure 2.14: Running time of the Combined Haplotype Test (CHT) with different numbers of individuals.

running time per site to be about 0.020 seconds per individual on Linux machines with Intel Xeon E5620 2.4 GHz and Intel Xeon L5420 2.5GHz CPUs (Figure 2.14).

2.8 Combined haplotype test caveats

WASP can only test for gene-level expression differences and does not consider the expression of individual transcript isoforms. Some QTLs detected by WASP may therefore be due to differences in isoform usage rather than differences in overall gene expression[46, 47].

2.9 Conclusions

Our results demonstrate that WASP is a powerful approach for the identification of molecular QTLs, particularly when sample sizes are small. WASP accounts for numerous biases in allele-specific data and is flexible enough to work with different read mappers and multiple types of sequencing data such as ChIP-seq and RNA-seq. By modeling biases and dispersion differences directly, WASP eliminates the need for quantile normalization of the data, thereby making estimated effect sizes more interpretable. The source code and documentation for WASP are open source and can be downloaded from https://github.com/bmvdgeijn/WASP/.

CHAPTER 3
IDENTIFICATION OF GENETIC VARIANTS THAT AFFECT HISTONE MODIFICATIONS IN HUMAN CELLS

3.1 Abstract

Histone modifications are important markers of function and chromatin state, yet the DNA sequence elements that direct them to specific genomic locations are poorly understood. Here we identify hundreds of quantitative trait loci, genome-wide, that impact histone modification or RNA polymerase (PolII) occupancy in Yoruba lymphoblastoid cell lines (LCLs). In many cases the same variant is associated with quantitative changes in multiple histone marks and PolII, as well as in DNaseI sensitivity and nucleosome positioning. Transcription factor binding site polymorphisms are correlated overall with differences in local histone modification and we identify specific transcription factors whose binding leads to histone modification in LCLs. Furthermore, variants that impact chromatin at distal regulatory sites frequently also direct changes in chromatin and gene expression at associated promoters.

Contribution

This work was done in collaboration with Graham McVicker. I did much of the statistical modeling, including the development of mark correlation test and extensions of the CHT. I made Figures 3.2, 3.6, 3.7, 3.9, 3.10, 3.12, 3.13.

3.2 Overview

Variation at noncoding regulatory sequences contributes to the genetics of complex traits [48, 49, 50], yet we still have limited understanding of the primary mechanisms by which they act. One possibility is that regulatory variants affect histone modifications that have downstream consequences on chromatin remodeling or transcription [51]. There are many possible post-translational modifications of histones (i.e., histone marks) [51], and sets of these co-occur in distinct chromatin states [52, 53, 54, 55, 56], are associated with functional elements [49, 57, 58], and are sensitive indicators of changes in gene regulation [56, 58]. However, we still do not know whether histone modifications are generally a cause or a consequence of gene regulation, or which DNA elements direct cell typeappropriate histone marking [54, 59]. Thus, studies of genetic variants that disrupt transcription factor binding sites may illuminate whether histone modifications enable transcription factor binding or whether the binding of transcription factors results in histone modification. We performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for RNA PolII and four post-translational modifications of histone H3 (H3K4me1, H3K4me3, H3K27ac and H3K27me3) in ten unrelated Yoruba LCLs. H3K4me3 (tri-methylation of lysine 4) is primarily associated with active promoters, H3K4me1 (mono-methylation of lysine 4) is associated with active chromatin outside of promoters (e.g. enhancers), H3K27ac (acetylation of lysine 27) is associated with both active promoters and enhancers [53, 60], and H3K27me3 (tri-methylation of lysine 27) is associated with silencing by the polycomb repressive complex 2 (PRC2) [61, 62].

3.3 Data generation and quality control

3.3.1 Samples and cell culture

Ten lymphoblastoid cell lines (LCLs) from unrelated Yoruba individuals were obtained from the Coriell Institute (Camden, NJ; http://www.coriell.org): GM18505, GM18507, GM18508, GM18516, GM18522, GM19141, GM19193, GM19204, GM19238, GM19239. The LCLs were grown in RPMI media with 15% FBS, supplemented with 2mM L-glutamate, 100 I.U./mL penicillin, and 100 μg/mL streptomycin.

3.3.2 ChIP-seq

ChIP-seq data were previously collected for H3K4me3 for three of the samples[63] and for PolII for six of the samples[17]. For the other samples, ChIP-seq was performed as described[63], except that chromatin was sheared with a Covaris (Woburn, MA) S2 (settings: 40 minutes, duty cycle 20%, intensity 8, 200 cycles/burst, 500 μL at a time in 12 × 24 mm tubes). We separately optimized the amount of antibody used for each type of experiment: H3K4me3 (4 μg, Abcam (Cambridge, MA) ab8580), H3K4me1 (12 μg, Millipore (Billerica, MA) 07-436), H3K27ac (4 μg, Abcam ab4729), H3K27me3 (4 μg, Millipore 07-449), and Pol II (10 μg, Santa Cruz Biotechnology (Dallas, TX) sc-9001).

The quality of each immuno-precipitation was assessed by RT-PCR of positive and negative control genomic regions that were previously shown to be enriched or not enriched for each datatype[48]. Successful ChIP assays showed enrichment at the positive control regions relative to the negative control regions in the immunoprecipitated sample (and compared to the input whole-cell extract from the same individual). We prepared Illumina (San Diego, CA) sequencing libraries from the DNA from each ChIP sample, and from a pooled input sample (containing equal amounts of DNA by mass) as previously described[48], starting with 20 μL of ChIP output or 4 ng of pooled input sample.

Libraries were sequenced in one or more lanes on an Illumina sequencing system using standard Illumina protocols. H3K4me3, H3K4me1, H3K27ac, and H3K27me3 samples were sequenced on a Genome Analyzer II (GAII) system (single end, 36 bp), and Pol II and input samples were sequenced on a HiSeq system (single end, 28 bp). Input reads were trimmed to 28 bp and 36 bp, where appropriate, for comparison to the data generated from ChIP samples.

3.3.3 Read mapping and sample validation

We mapped sequence reads to the human reference genome (hg18) using BWA[32], allowing up to 2 mismatches per read (`-n 2`), and excluding gapped alignments (`-o 0`). Total reads and mapping statistics for each individual and datatype are given in Table 3.3.3.

Table 3.1: Total sequenced, uniquely mappable and non-duplicate mapped reads for each sample.

Datatype	Individual	Total reads	Mappable reads	Non-duplicate reads
H3K27ac	18505	41,619,485	34,819,891	33,388,658
H3K27ac	18507	36,534,335	31,176,526	30,297,842
H3K27ac	18508	42,590,850	36,133,311	34,859,856
H3K27ac	18516	31,212,054	26,976,260	25,413,627
H3K27ac	18522	41,753,448	35,269,619	34,205,704
H3K27ac	19141	36,602,602	31,630,022	30,691,817
H3K27ac	19193	42,554,122	35,212,025	34,091,012
H3K27ac	19204	38,707,056	32,493,582	31,020,227
H3K27ac	19238	42,518,152	35,794,207	34,863,680
H3K27ac	19239	42,985,514	36,343,931	34,694,833
H3K27me3	18505	42,418,783	35,394,414	32,360,537
H3K27me3	18507	41,882,091	34,050,036	31,288,003
H3K27me3	18508	43,683,243	36,672,935	33,783,533
H3K27me3	18516	85,601,902	73,883,725	52,791,050
H3K27me3	18522	41,465,617	34,031,717	27,271,414
H3K27me3	19141	41,396,087	33,595,112	29,518,442
H3K27me3	19193	42,376,103	34,519,855	32,530,869
H3K27me3	19204	41,769,273	34,216,184	32,573,801
H3K27me3	19238	41,259,657	34,436,378	31,794,352
H3K27me3	19239	40,014,399	32,115,822	27,258,931
H3K4me1	18505	55,852,093	46,027,876	21,588,975
H3K4me1	18507	82,589,169	67,136,686	31,143,373
H3K4me1	18508	16,648,997	13,489,819	8,704,990
H3K4me1	18516	42,389,962	32,311,347	13,652,759
H3K4me1	18522	29,251,590	23,565,863	16,590,932

Table 3.1, continued

Datatype	Individual	Total reads	Mappable reads	Non-duplicate reads
H3K4me1	19141	24,705,715	19,895,574	7,136,047
H3K4me1	19193	33,200,484	27,376,219	19,907,413
H3K4me1	19204	39,385,609	32,476,133	18,568,012
H3K4me1	19238	33,585,084	27,992,337	12,371,015
H3K4me1	19239	41,487,566	34,107,283	26,942,630
H3K4me3	18505	42,386,132	36,134,558	33,392,023
H3K4me3	18507	41,163,781	33,100,647	28,724,717
H3K4me3	18508	39,410,115	34,226,880	30,650,658
H3K4me3	18516	33,418,845	26,707,043	24,195,825
H3K4me3	18522	42,253,530	35,864,632	22,021,929
H3K4me3	19141	35,600,431	29,164,849	22,436,825
H3K4me3	19193	33,272,920	26,786,692	24,517,535
H3K4me3	19204	29,135,614	22,934,303	21,236,394
H3K4me3	19238	31,528,815	25,722,058	18,636,764
H3K4me3	19239	41,049,455	33,789,330	26,789,009
PolII	18505	409,314,862	312,809,309	33,182,679
PolII	18507	207,080,008	157,306,899	27,401,762
PolII	18508	206,275,097	158,488,158	38,943,864
PolII	18516	374,652,279	289,734,350	41,269,542
PolII	18522	206,977,123	158,020,656	52,161,952
PolII	19141	202,946,440	155,229,617	52,303,705
PolII	19193	167,057,685	123,473,127	48,793,162
PolII	19204	203,922,597	148,352,038	59,454,889
PolII	19238	598,313,128	452,430,530	123,773,435
PolII	19239	204,235,178	157,764,110	58,732,355

3.3.4 Heirarchical clustering to confirm data

We used hierarchical clustering to verify that the general properties of each library were consistent with those from the same chromatin feature in the ENCODE dataset and the other libraries in our dataset. For each lane of sequencing data, we extracted read counts within a 2 kb window of each annotated Ensembl transcription start site. Read counts were quantile normalized and hierarchical clustering was performed on a matrix of Pearson correlations between all pairs of quantile-normalized counts. All sequencing lanes from each distinct ChIP experiment-type formed non-overlapping clusters, and these clusters included the corresponding ChIP experiments from the ENCODE project, with the exception of two lanes of data labeled in the ENCODE project as H3K27ac, which appear to be of poor quality (Figure 3.1). To check for contamination among cell lines and mislabeling of samples during processing, the reads from each library were checked for consistency with published genotypes[12]. All libraries could be confidently assigned to a single individual and were retained for further analysis.

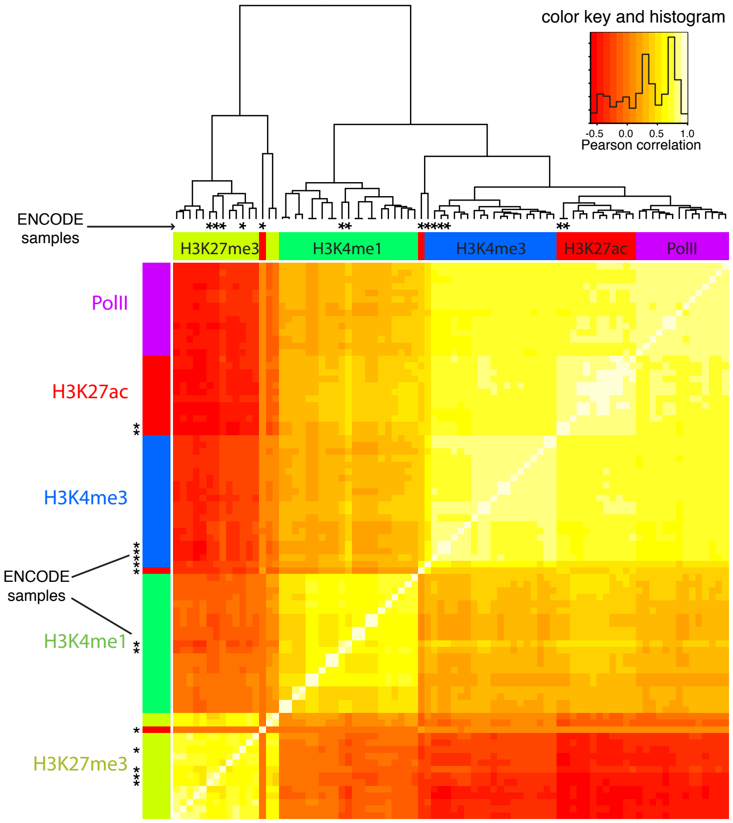

Figure 3.1: **Clustering of ChIP-seq data from ENCODE and this study.**. The heatmap shows hierarchically clustered ChIP-seq data, using pairwise Pearson correlation as a distance metric. Correlations were calculated from quantile-normalized read counts from each flowcell lane. Read counts were extracted from 2 kb windows centered on annotated Ensembl transcription start sites. Colored bars beside the heatmap indicate the datatype label of each sample. ENCODE samples are indicated with (*); the other samples were collected for this study.

3.3.5 Controlling for allelic differences in mappability

Sequence polymorphisms can cause substantial mapping biases and false allele-specific signals[27]. To control for mapping biases we used a custom read mapper that reports the uniqueness of reads originating from each genomic position, while taking into account sequence polymorphisms[22]. We discarded all reads that the mapper reported as non-uniquely mapping. This mapper only considers the first 20 bp of each read (due to memory constraints), so in most cases its estimates of mapping uniqueness are conservative for our 28 bp and 36 bp reads. One issue, however, is that reads can incorrectly be reported as uniquely mapping when multiple polymorphisms occur in close proximity to one another (greater than 20 bp apart, but less than or equal to 36 bp apart). To account for this problem, we additionally filtered all mapped reads that overlapped more than one polymorphism.

3.3.6 Filtering duplicate reads

When multiple reads from the same sample mapped to the same genomic location, we discarded all but one to avoid artefacts caused by PCR and optical duplicates. Duplicates were discarded randomly rather than taking the highest scoring reads, because the latter approach is biased towards keeping reads that match the reference genome.

3.3.7 Genotype imputation and phasing

We imputed genotypes and phased our samples with IMPUTE2[35] using the 1000 Genomes Phase1 integrated version 3 reference panel[12]. To speed up computation, we used pre-phasing information[64] from HapMap Phase II genotypes (release 22)[21]. We used the IMPUTE2 option `-filt_rules_l 'afr.maf$<$0.004'` to remove sites that are monomorphic or singletons in the 246 AFR individuals in the 1000 Genomes panel and the `-Ne 20000` option to specify an effective population size of 20,000. Since the 1000 Genomes reference panel is on the hg19 assembly, we used liftover[65] to transfer HapMap genotype and phase information from hg18 to hg19. We removed SNPs with strands, chromosomes or ordering that differed between hg18 and hg19. After imputation, we transferred the SNPs back to hg18 using liftover.

3.3.8 RNA-seq, DNaseI-seq and MNase-seq data

For plotting RNA-seq read depths, we obtained RNA-seq reads from 69 unrelated Yoruba LCLs[15] and mapped them to the human reference genome (hg19) using BWA[32]. We excluded read alignments with gaps, more than 2 mismatches, or mapping quality scores less than 10. We computed read depth at each position by summing overlapping reads, and converted coordinates to hg18 using a custom script.

RNA-seq expression measurements for Ensembl genes and eQTL calls were previously calculated by our lab[22]. Mapped DNaseI-seq reads and dsQTL calls from 70 unrelated Yoruba LCLs were obtained from the same study[22].

Nucleosome dyad positions from mapped MNase-seq reads for 7 unrelated Yoruba LCLs were previously collected in our lab[7].

The RNA-seq, DNase-seq, and MNase-seq data are available from GEO (www.ncbi.nlm.nih.gov/geo/) under accessions GSE19480, GSE31388, and GSE36979.

3.4 Mapping histone modification QTLs

To identify genetic associations with histone marks and PolII, we used an early version of the WASP combined haplotype test, which is described in Chapter 2. The following features were not included in this version:

- Maximum likelihood estimation of dispersion parameters from the data,
- Adjustment for unknown covariates by allowing principal component loadings to be provided,
- Allowing tested regions to be split across multiple genomic segments, such as exons,

- Greater efficiency so the model can be run with hundreds of individuals.

We applied the combined haplotype test to hundreds of thousands of polymorphic sites with sufficient read depth (i.e., sites within ChIP-seq peaks) and identified over 1,200 histone mark and PolII QTLs at a false discovery rate (FDR) of 20% (Figure 3.2). After merging overlapping regions, we identified a total of 27 distinct QTLs for H3K4me1, 469 for H3K4me3, 730 for H3K27ac, 118 for PolII, and 2 for H3K27me3 (which tends not to fall into strong peaks) (Table 3.2). At an FDR threshold of 10% we identified 582 distinct histone mark and PolII QTLs (Table 3.2). In principle some of these signals might be due to imprinting or random allelic inactivation; however, several lines of evidence indicate that most of the regions we identify are conventional QTLs (see later section).

Table 3.2: **Summary of results from the genome-wide combined haplotype test.** For each datatype, the columns provide the number of SNPs tested; the number of significant SNPs at different false discovery rate (FDR) thresholds; and the number of distinct significant regions after merging those that overlap. The counts in the combined row are from the union of all datatypes. The small overlap in significant regions across datatypes is likely because there is poor power to identify overlapping QTLs by testing each datatype independently [66].

		Significant SNPs			Merged significant regions		
Datatype	Tested SNPs	fdr10%	fdr20%	fdr50%	fdr10%	fdr20%	fdr50%
H3K4me1	46,257	19	57	563	8	27	289
H3K4me3	111,732	741	1,219	3,444	246	469	1,527
H3K27ac	217,737	886	1,699	6,546	335	730	3,043
H3K27me3	233,604	4	4	9	2	2	5
PolII	412,406	173	303	759	58	118	381
combined		1,633	2,975	10,544	582	1,232	4,768

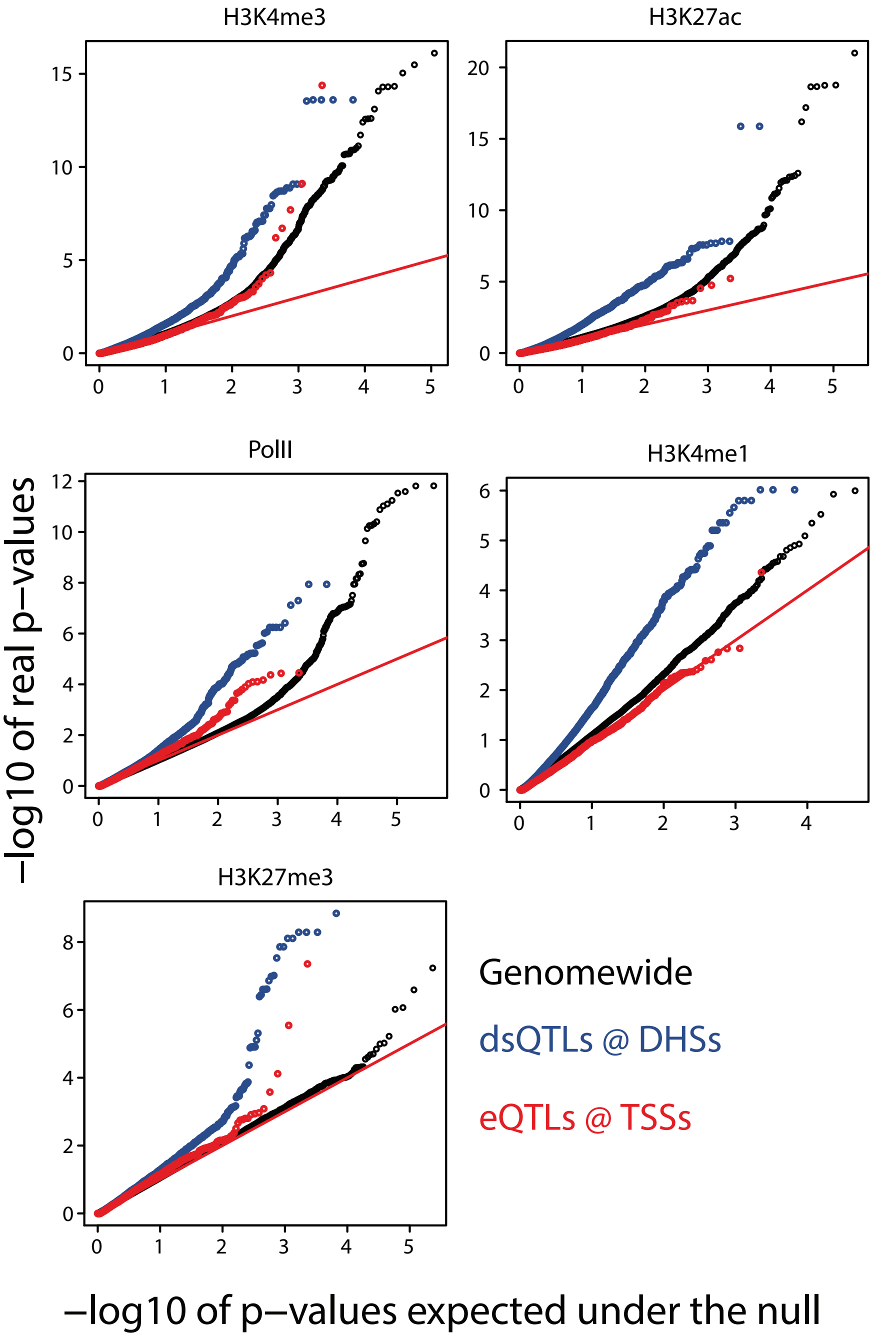

Figure 3.2: **Combined haplotype test results.** Quantile-quantile plots comparing -log10 p-values expected under the null to those from the combined test for association between genotype and allelic imbalance for H3K4me3, H3K27ac, RNA Polymerase II, H3K4me1, and H3K27me3. We applied the combined test to regions surrounding ChIP-seq peaks (Genomewide), to DNaseI hypersensitive sites (DHSs) associated with dsQTLs, and to transcription start sites (TSSs) associated with eQTLs.

3.4.1 Details on applying the combined test genome-wide

For each mark, we extracted total read depth and allele-specific counts in a 2 kb window around every SNP that was segregating in our sample. A site was considered testable in the genome-wide test if there were at least 15 informative reads overlapping heterozygous SNPs. By this measure, there were between 44,000 and 415,000 testable sites for each mark (Table 3.2). We applied our combined haplotype test to each of these sites and identified significant associations at an FDR of 10% or 20%[67] (Figure 3.2). Overlapping windows were then merged to get a count of unique associated sites.

3.4.2 Permutations to assess calibration

To assess the calibration of the test, we also applied it to permuted data. We permuted total read depths (and matching genome-wide read depth counts) and randomly flipped allele-specific counts at linked heterozygous SNPs with probability 0.5. The permuted results showed little to no signal, so we conclude that our test is well calibrated. It is possible that non-genetic monoallelic inactivation could cause signal in the allele specific part of the test even if there is no genetic determinant. We therefore did a second version of the permutations which maintained the haplotype information. At each linked region, we either flipped the allele specific counts at every SNP or none of them with probability 0.5 so that any monoallelic effect would be maintained. These permuted results also showed greatly reduced signal so we are confident that most of our observed signal reflects true genetic associations.

3.5 Evidence that histone mark and PolII QTLs are real QTLs

Several lines of evidence indicate that most of the histone mark and PolII QTLs that we identify are real and not due to random allelic inactivation, imprinting or technical artifacts. First, we were very careful to remove all sources of read mapping bias, a well-known source of false-positives[27]. Second we incorporate over-dispersion into our statistical models, which accommodates unknown sources of non-genetic variation. Third, histone mark and PolII QTLs are enriched at dsQTLs and eQTLs (Figure 3.2), which were identified with large sample sizes and without allele-specific information [22, 15]. Fourth, we see a dearth of opposite direction effects between the activating histone mark QTLs, PolII QTLs, dsQTLs and eQTLs (Figures 3.6,3.7). If our new QTLs were due to non-genetic factors, same- and opposite-direction effects should occur with similar frequencies. Fifth, permuting entire haplotypes (while maintaining phase of alleles) removes most enrichment of low p-values from the combined haplotype test which is not expected if the results were due to random allelic inactivation. Together, these observations argue that most of the significant regions we identify are true genetic associations.

3.6 Histone modification QTLs overlap previously identified QTLs

Many of the histone mark QTLs overlap previously identified QTLs for DNaseI sensitivity (denoted dsQTLs) [22]. DNaseI sensitivity is an indicator of open chromatin, and DNaseI hypersensitive sites (DHSs) typically mark active regulatory regions that are associated with active histone marks and transcription factor binding [68]. Indeed, we found an enrichment of low p-values when testing for QTL associations with PolII and all four histone marks at dsQTLs, compared to the genome-wide set of tested single nucleotide polymorphisms (SNPs) (Figure 3.2).

3.6.1 Distance between tested SNPs and dsQTL DHSs or eQTL TSSs

We calculated the distances between the SNPs that were tested in the genome-wide combined test and the nearest previously-identified dsQTL DNase hypersensitive sites (DHSs) and eQTL transcription start sites (TSSs)[22]. We also calculated distances for the subset of SNPs that were significant at a false discovery rate of 20% (when multiple significant regions overlapped only the most significant SNP was used) (Figures 3.3&3.4). QTLs for histone marks and PolII are significantly enriched (compared to tested SNPs) near dsQTL DHSs, although a substantial fraction of them are found further away (Table 3.3). This suggests that histone modifications may provide more power to detect differences in chromatin state beyond that of DNaseI sensitivity.

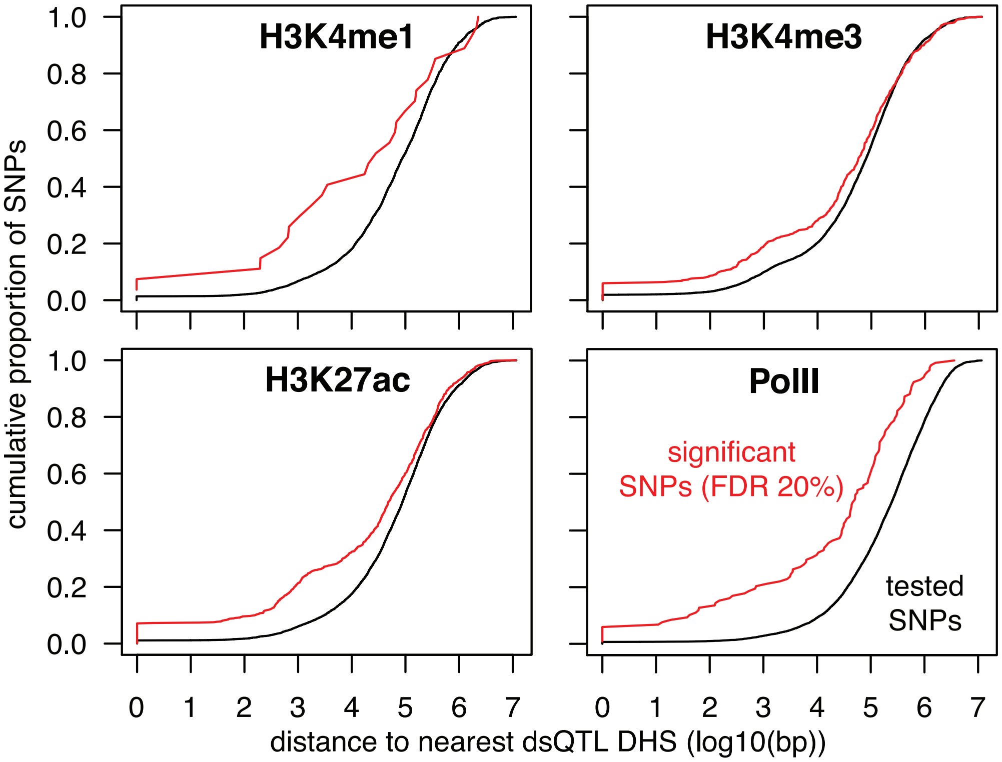

Figure 3.3: Cumulative distributions of distances between tested SNPs and dsQTL DHSs

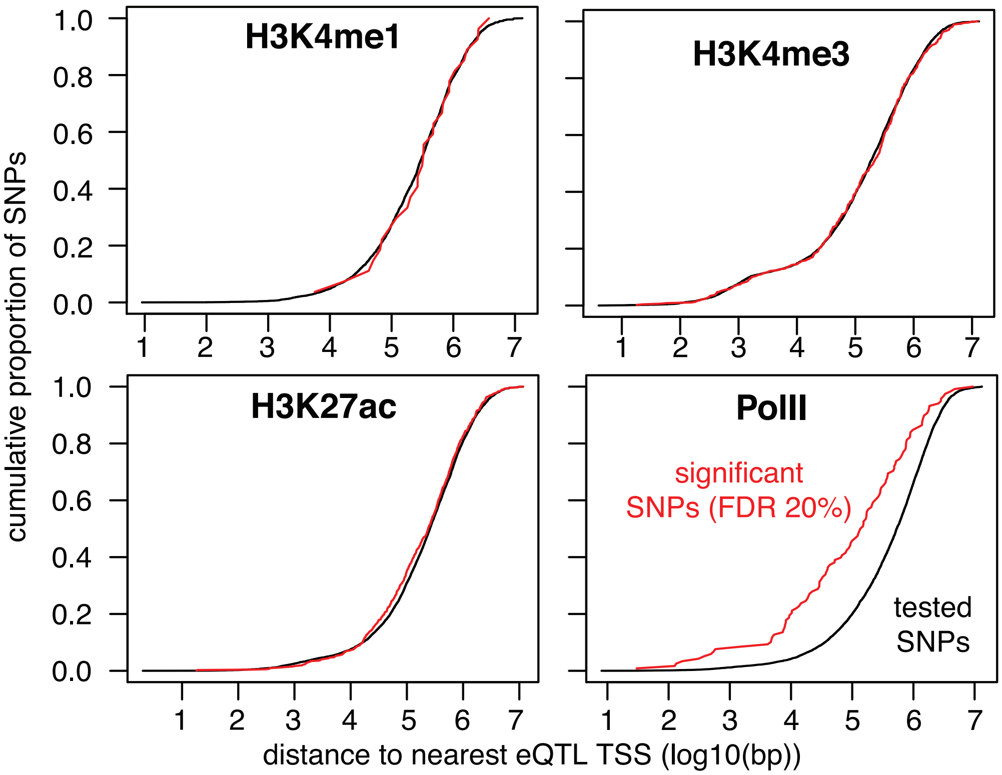

Figure 3.4: Cumulative distributions of distances between tested SNPs and eQTL TSSs

Table 3.3: **Enrichment of histone mark and PolII QTLs near dsQTL DHSs** "All SNPs" are a random subset (10,000) of the complete set of tested SNPs that are matched for minor allele frequency with the significant SNPs. "Signif. SNPs" are those that are significant at an FDR threshold of 20%. SNPs are considered "near" a dsQTL DHS if they are within 1kb, and "far" otherwise. The odds ratio is the ratio of near to far significant SNPs divided by same ratio for all SNPs. The p-value is from a two-sided Fisher's Exact Test with the alternative hypothesis that the true odds ratio is not equal to 1.0.

	All SNPs		Signif. SNPs			
Datatype	Near DHS	Far DHS	Near DHS	Far DHS	Odds Ratio	p-value
H3K4me1	684	9315	7	20	4.77	2×10^{-3}
H3K4me3	994	9005	90	379	2.15	4×10^{-9}
H3K27ac	607	9392	156	574	4.21	1×10^{-38}
PolII	280	9719	24	94	8.86	6×10^{-14}

Table 3.4: **Enrichment of histone mark and PolII QTLs near eQTL TSSs** The columns of this table are as described for Table 3.3, but are computed for distance from eQTL TSSs rather than dsQTL DHSs

	All SNPs		Signif. SNPs			
Datatype	Near TSS	Far TSS	Near TSS	Far TSS	Odds Ratio	p-value
H3K4me1	50	9949	0	27	0	1.0
H3K4me3	774	9225	34	435	0.932	0.79
H3K27ac	250	9749	12	718	0.652	0.17
PolII	117	9882	9	109	6.97	2×10^{-5}

Table 3.5: **Genomic locations of histone mark and PolII QTLs** This table gives the numbers of QTLs (at FDR 20%) that are within 1 kb of (or within) a DNase hypersensitive site or an annotated transcript. H3K27me3 is omitted because of the small number of QTLs for this modification. DHSs were identified by taking the top 1% of sites in the genome after smoothing aggregate DNase-seq read counts with a 100 bp sliding window.

Datatype	Total QTLs	$< 1kb$ from DHS	$< 1kb$ from transcript	$< 1kb$ from DHS or transcript
H3K4me1	27	8	8	15
H3K4me3	469	320	286	395
H3K27ac	730	381	407	557
PolII	118	79	65	95

3.7 Expression QTLs and DNase QTLs show aggregate effects on histone modifications

We plotted aggregate ChIP-seq read depth around DHSs associated with dsQTLs (Figure 3.5), grouping read counts according to whether an individual carries the genotype associated with high, medium or low sensitivity at a dsQTL. Most of the dsQTLs lie outside promoters, and the average histone mark read depths at dsQTL DHSs follow qualitative expectations for distal enhancers [53], with higher levels of H3K4me1, and lower levels of H3K4me3 and PolII compared to promoters. High sensitivity genotypes tend to have reduced nucleosome occupancy within the DHS (20); higher levels of transcription factor binding [22]; higher levels of the active marks H3K4me1, H3K4me3 and H3K27ac; and higher PolII occupancy. The relationship between DNaseI and the repressive mark H3K27me3 is more complicated, as we find both positive and negative associations. We find no opposite-direction effects between DNaseI and either H3K4me1, H3K4me3 or H3K27ac (Figure 3.6). At expression QTLs (eQTLs) [15], we stratified the samples by the genotype of the most significant eQTL SNP, and found overall patterns similar to those at dsQTLs (Figure 3.5). Individuals who are homozygous for the high-expression genotype generally have higher levels of DNaseI sensitivity, H3K4me3, H3K27ac and PolII occupancy [69] at transcription start sites (TSSs). The repressive H3K27me3 mark shows the opposite trend and is highest in the low-expression genotype class.

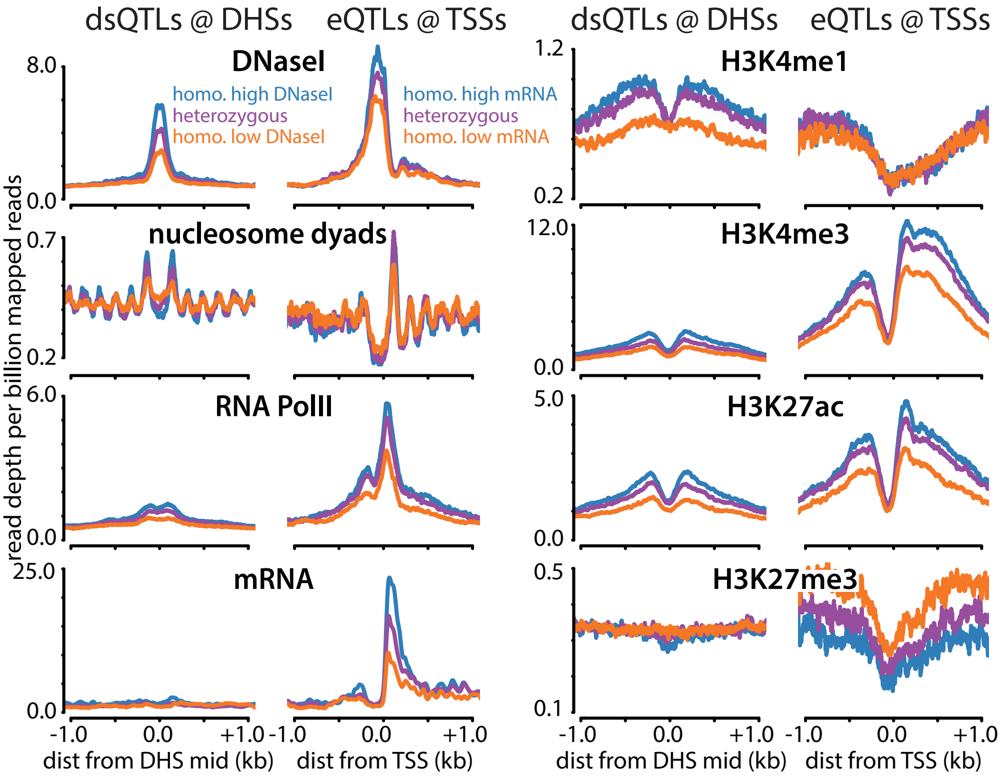

Figure 3.5: **Multiple molecular phenotypes are associated with the same genetic variants.** Panels show aggregate read depth for molecular traits at DNaseI hypersensitive sites (DHSs) associated with dsQTLs, or transcription start sites (TSSs) associated with eQTLs. Reads are grouped into high, medium and low sensitivity genotypes for dsQTLs; and high, medium and low expression genotypes for eQTLs. Plots were made from half of the significant dsQTLs and eQTLs (those with the lowest p-values; n=2787 for dsQTLs; n=638 for eQTLs).

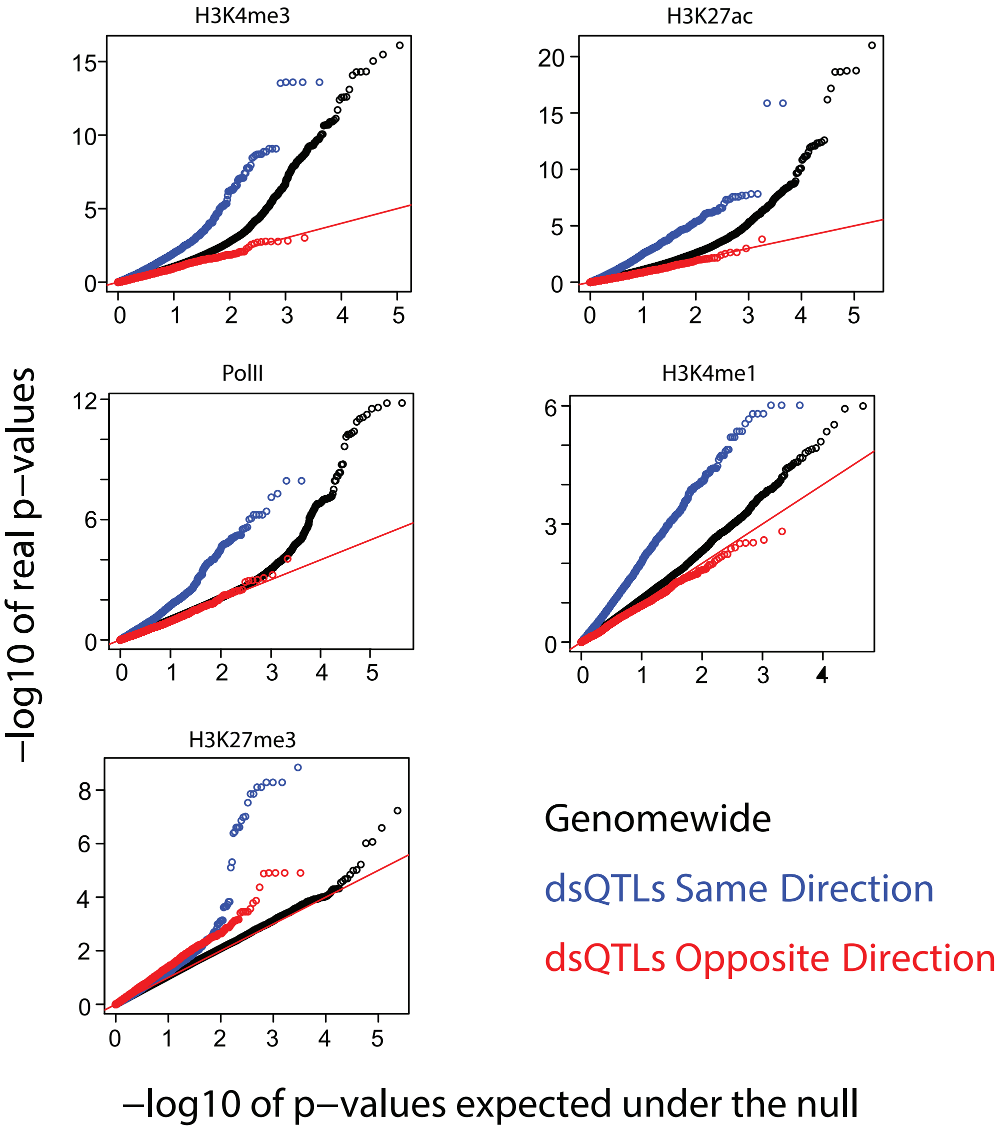

Figure 3.6: **Polarized effects of dsQTLs on marks at DHS regions.** Quantile-quantile plots comparing -log10 p-values expected under the null to those from the combined test applied to DNaseI hypersensitive sites (DHSs) associated with dsQTLs. Regions were stratified by whether their estimated effects were in the same or opposite direction as the change in DNaseI sensitivity. Effects were considered to be in the same direction if the high sensitivity allele was associated with increased marking.

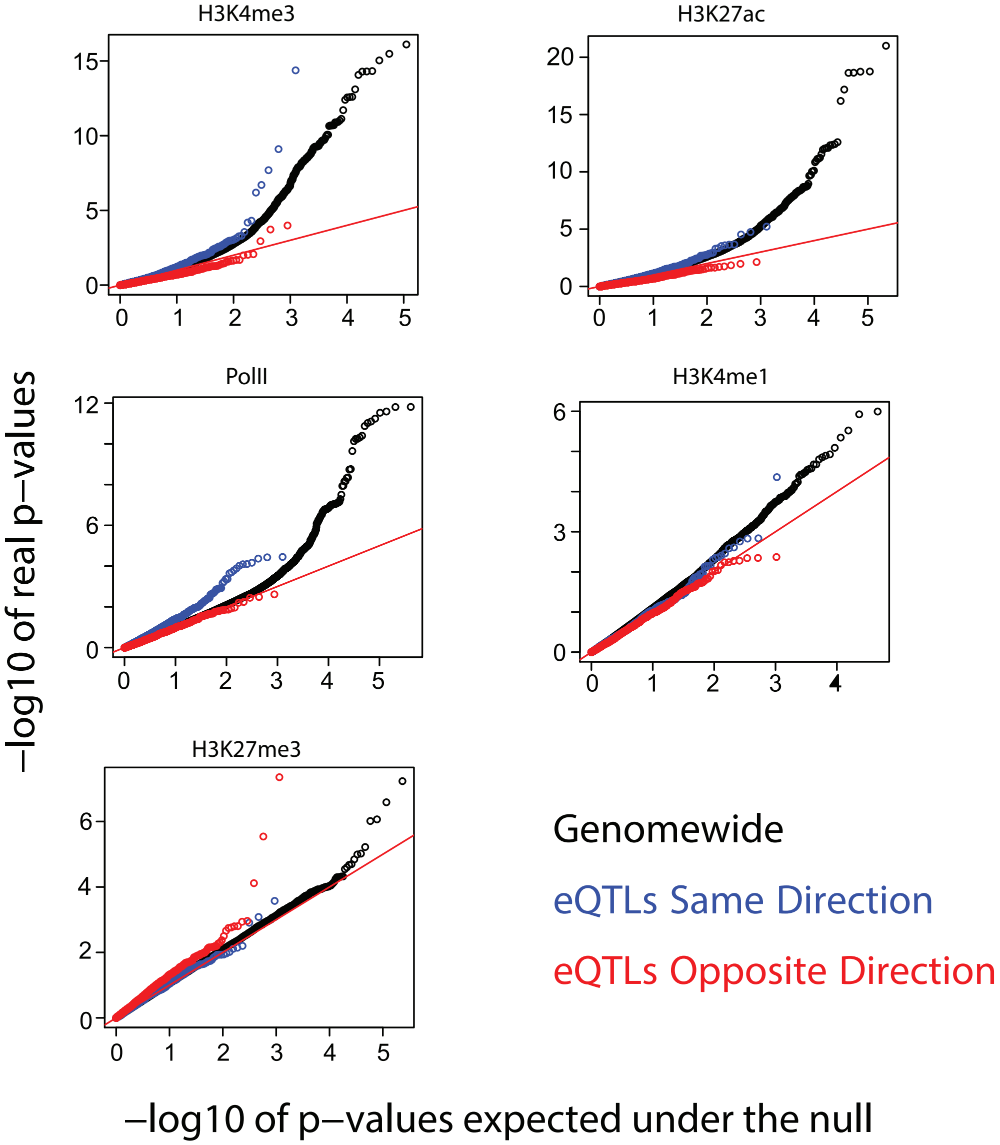

Figure 3.7: **Polarized effects of eQTLs on marks around TSSs.** Quantile-quantile plots comparing -log10 p-values expected under the null to those from the combined test applied to transcription start sites (TSSs) of genes associated with eQTLs. eQTLs were stratified by whether their estimated effect was in the same or opposite direction as the change in gene expression. Effects were considered to be in the same direction if the high expression allele was associated with increased marking.

3.8 QTL changes are often coordinated across phenotypes

When visualizing our top QTL hits for each of our histone marks, we noticed that many of the QTLs were shared across measurements. Indeed, it was often the case that a single SNP is associated with changes in multiple histone modifications, DNase sensitivity, and expression at a nearby gene. (Figure 3.8)

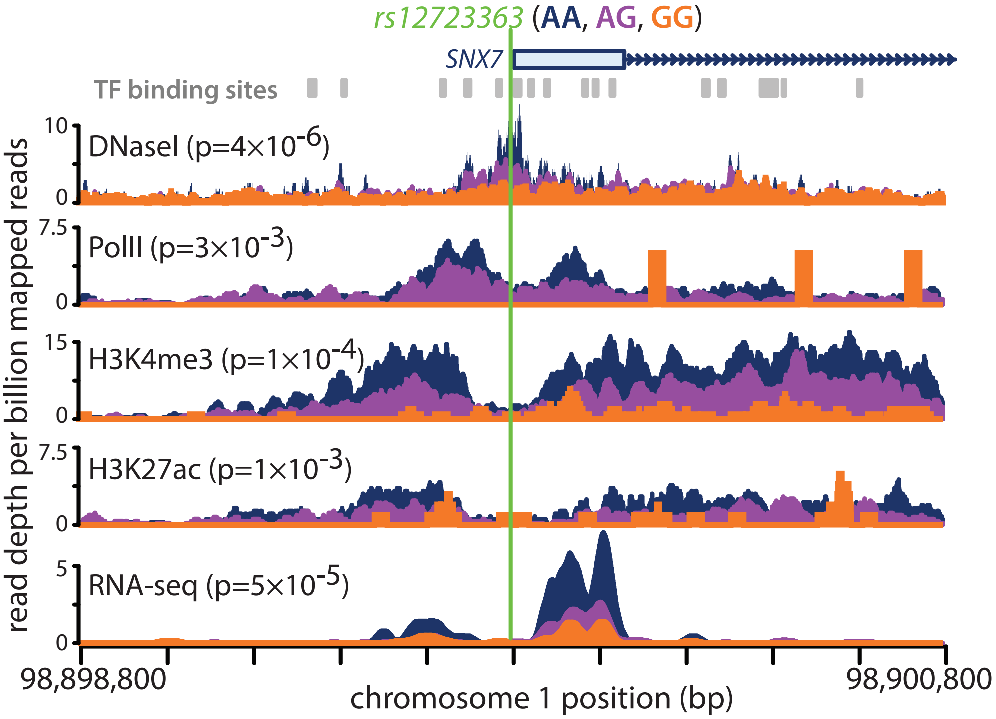

Figure 3.8: **An example of a QTL for multiple molecular phenotypes including DNaseI sensitivity, gene expression, H3K4me3, H3K27ac and PolII levels.** The tracks are colored by the genotype of the SNP rs12723363. P-values were computed with the combined haplotype test, except for DNaseI and RNA-seq where a linear model (t-test) was used.

3.9 Allelic imbalance is correlated across histone marks, PolII and DNaseI

We estimated the correlation of allele-specific changes across pairs of data types, while accounting for the sampling variance at individual sites. The allelic imbalances for features associated with active regionsDNaseI, PolII, H3K4me1, H3K4me3 and H3K27acare all highly positively correlated across 2 kb windows centered at dsQTL DHSs (Figure 3.9). In particular, the strong correlation in H3K4me3 and H3K27ac allelic imbalances indicates that these modifications are functionally linked and often depend on the same genetic elements.

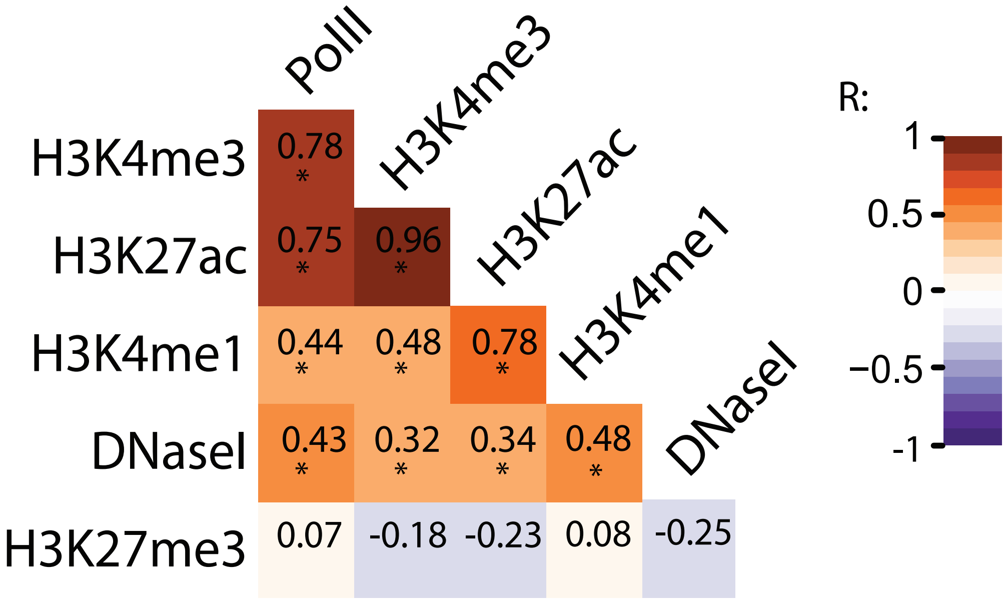

Figure 3.9: **Correlation in allelic imbalance between data types.** Correlation in allelic imbalance between data types at dsQTLs (* indicates $p < 10^{-3}$ by likelihood ratio test).

3.9.1 Model details

To investigate correlation between allele-specific differences in pairs of marks, we developed a method to estimate the covariance of allelic imbalances while accounting for variance due to limited read depth. We consider each region h to have an underlying allelic imbalance p_h, and we want to test whether the p_hs correlate across marks. We assume that when each mark is considered separately, $\operatorname{logit}(p_h)$ is distributed normally with mean $\mu = 0$ and variance σ^2.

$$\log\left(\frac{P}{1-P}\right) \sim \text{Norm}\left(\mu = 0, \sigma^2\right)$$

If the variances are relatively small, the p_hs will also be distributed approximately normally with $\mu = 0.5$, however the logit scale is more flexible in that it can handle cases where the variance is larger and most of the p_hs are close to 0 or 1.

When considered jointly, the allelic imbalances for two marks are assumed to be distributed as multivariate normal with an extra parameter, ρ, which describes their correlation.

$$\left(\log\left(\frac{P_1}{1-P_1}\right), \log\left(\frac{P_2}{1-P_2}\right)\right) \sim \text{MVNorm}\left(\mu, \Sigma\right)$$

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Given a p_h, the allele-specific counts for individual i at linked SNP j in region h are binomially distributed:

$$X_{i,j,h} \sim \text{Binom}\left(p_h, n_{i,j,h}\right)$$

This gives the following likelihood equation for the data, D, which consists of allele-specific read counts, $x_{1,i,j,h}$ and $x_{2,i,j,h}$ for marks 1 and 2:

$$\mathrm{L}\left(D\left|\sigma_1,\sigma_2,\rho\right.\right)=\prod_{h\in\text{regions}}\int_0^1\int_0^1\text{MVNorm}\left(\text{logit}\left(p_1\right),\text{logit}\left(p_2\right)\left|\sigma_1,\sigma_2,\rho\right.\right)$$

$$\prod_{i\in\text{inds}}\prod_{j\in\text{linkSNPs}}\Pr_{\text{Bin}}\left(x_{1,i,j,h}\left|p_1,n_{1,i,j,h}\right.\right)\Pr_{\text{Bin}}\left(x_{2,i,j,h}\left|p_2,n_{2,i,j,h}\right.\right)dp_1\,dp_2$$

It is computationally slow to evaluate the double integral numerically, particularly when it must be done many times for the likelihood maximization process. We instead obtained an analytic approximation to the double integral using a Laplace transformation[70]. This allows us to efficiently calculate maximum likelihood estimates of σ_1, σ_2, and ρ.

3.9.2 Applying the Model

We extracted the read counts for each mark from 2 kb windows centered on DNaseI hypersensitive sites (DHSs) associated with dsQTLs. Reads overlapping phased heterozygous SNPs were assigned to each haplotype and we estimated the correlation in allelic imbalance between pairs of marks, ρ, using maximum likelihood. Significance was assessed by comparing to a null model of no correlation $\rho = 0$, using a likelihood ratio test. We measured correlations in allelic imbalance for all pairs of histone marks and PolII. We also estimated correlations in allelic imbalance with DNaseI, but because DNaseI tends to be much more sharply peaked, we extracted DNaseI reads from a smaller region (200 bp).

3.9.3 Alternative method using harmonic weighted regression

To verify the results of our model, we also estimated correlation in allelic imbalance using a simpler method. We used the proportion of reads from haplotype 1 at heterozygous SNPs in each region (combining across individuals and linked SNPs) as an estimate of $p_{h,\bullet,\bullet}$, and performed linear regression of the $\hat{p}_h$s from one mark versus another, weighting the regression by the harmonic mean of the reads from each genotype:

$$w_h = \frac{2}{\frac{1}{n_{1,h,\bullet,\bullet}} + \frac{1}{n_{2,h,\bullet,\bullet}}}$$

The resulting pairwise correlations of allelic imbalances from this method were very similar to those from the other model. Most of the correlations were still significant, their signs remained the same, and the relative magnitudes of correlated mark pairs were conserved between tests. The absolute magnitudes of the correlations were considerably smaller because this method does not account for variation introduced by binomial sampling.

3.10 SNPs that are dsQTL-eQTLs consistently affect enhancer and promotor modifications

Since dsQTLs are frequently also eQTLs [22], we used dsQTLs that are eQTLs (dsQTL-eQTLs) to assign DHSs to TSSs. We classified dsQTL-eQTLs as activating if the high DNaseI sensitivity allele was also the high gene expression allele, and as repressing otherwise (Figure 3.10A). We confirmed that most activating dsQTL-eQTLs are true joint associations (as opposed to independent QTLs in linkage disequilibrium), but discarded the repressive dsQTLs because only a small number had lower p-values than expected by chance (fig. S10). We only used dsQTL-eQTLs where the associated DHS was at least 5 kb away from the associated TSS so the regions are likely to be functionally distinct. For each dsQTL-eQTL pair, we estimated average allelic imbalance in histone marks and PolII after polarizing genotypes by DNaseI sensitivity at the associated DHS. At activating DHSs, the allelic imbalance is positive (in the same direction as DNaseI sensitivity at the DHS) for the three activating histone marks and PolII, and is negative for H3K27me3 (Figure 3.10A). The same pattern is present at the associated TSSs, which demonstrates that polymorphisms can jointly affect chromatin state at distal enhancers and at promoters, perhaps via chromatin looping interactions [69]. We found that for several of the dsQTL-eQTLs, the SNP that is most significantly associated with DNaseI sensitivity is located in a binding site for a known transcription factor (Figure 3.10B).

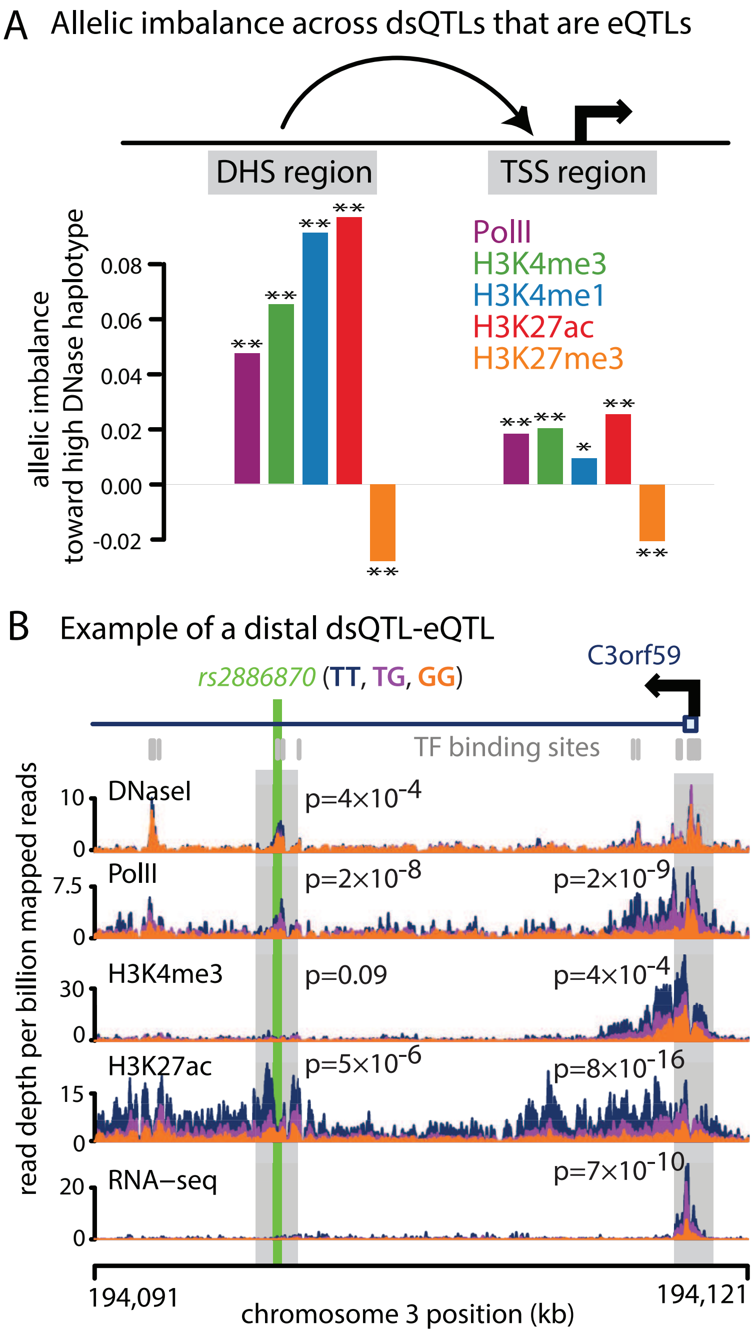

Figure 3.10: **Histone modification changes at dsQTLs that are also eQTLs.** (A) Estimates of allelic imbalance for histone marks and PolII across DNaseI hypersensitive sites (DHSs; n=239) and transcription start sites (TSSs; n=246) from joint dsQTL-eQTLs (17). (*) and (**) indicate allelic imbalance is significantly different from 0 with $p < 0.05$ and $p < 0.01$, respectively (by likelihood ratio test). (B) An example of a dsQTL-eQTL. The SNP rs2886870 disrupts an NF-B binding site and is significantly associated with local DNase sensitivity, H3K27ac and PolII levels. The SNP is also significantly associated with gene expression, H3K27ac, H3K4me3 and PolII levels at a distal promoter (>18 kb away). P-values are from the combined haplotype test, except for DNaseI and RNA-seq where linear regression and t-tests were used. Read depth tracks are aggregated and colored by the genotype of rs2886870.

3.10.1 Extending the CHT to multiple sites

Our model for measuring the aggregate allelic imbalance is similar to the combined haplotype test that we apply to a single region at a time. A few key modifications make it easier to apply the test across many regions at once. At each site h, instead of modeling the read depth for each individual as a Poisson distribution, we model the distribution of the reads between individuals given the total read depth across all individuals using a multinomial distribution:

$$\left(X_{h,1}...X_{h,10}\right) \sim \text{Multinomial}\left(\rho, T_{h,\bullet}\right)$$

$$\rho_{ij} = \begin{cases} 2pT_{h,i}/C & \text{if } G_{h,i} = 0 \text{ (homozygote high DNaseI)} \\ \\ T_{h,i}/C & \text{if } G_{h,i} = 1 \text{ (heterozygote)} \\ \\ 2(1-p)T_{h,i}/C & \text{if } G_{h,i} = 2 \text{ (homozygote low DNaseI)} \end{cases}$$

By removing α and β from the equation, we eliminate the need to estimate the relative levels of marks at each individual site and thereby greatly reduce the number of parameters that need to be estimated. This proves less powerful when applied to one site at a time, but is crucial for combining across sites. We found that our data were overdispersed when we tried to model them with the Multinomial distribution. We therefore introduced an extra dispersion parameter, Ψ , which we estimate across all sites. This makes the distribution Dirichlet-Multinomial:

$$\mathbf{X_{h,*}} \sim \text{DMN}\left(\rho, \Psi, T_{h,\bullet}\right)$$

$$Pr_{\text{DMN}}\left(\mathbf{X} = x_{h,1}..x_{h,n}\left|p, \Psi\right.\right) = \frac{\Gamma\left(A_h\right)}{\Gamma\left(N_h + A_h\right)} \prod_{k \in 1..n} \frac{\Gamma\left(n_{h,k} + \alpha_{h,k}\right)}{\Gamma\left(\alpha_{h,k}\right)}$$

$$\alpha_{h,j} = \rho_{h,j}\Psi$$

$$N_h = \sum_{k \in 1..n} n_{h,k}$$

$$A_h = \sum_{k \in 1..n} \alpha_{h,k} = \Psi$$

A likelihood ratio test can then be conducted using the new likelihood equation:

$$\mathrm{L}\left(D\,|p,\Psi,\Upsilon\right) = \prod_{h} \Pr_{\mathrm{DMN}}\left(\mathbf{X_{h,\bullet}}\,|p,\Psi\right) \prod_{i}\prod_{j} \Pr_{\mathrm{BB}}\left(Y_{h,i,j} = a_{h,i,j}\,|p, n_{h,i,j}, \Upsilon\right)$$

3.10.2 Identifying dsQTL-eQTLs

We identified dsQTL-eQTLs using gene expression and DNaseI sensitivity data that were previously generated and processed by our group[22]. We started with a set of 6070 dsQTLs[22] that were within 100 kb of an Ensembl-annotated TSS, and tested each of the dsQTL SNPs for association with gene expression by regressing the normalized expression level of each individual against the number of copies of the non-reference allele that they carry. We classified putative dsQTL-eQTLs as activating if the high expression allele for the eQTL was also the high DNaseI sensitivity allele, and as repressing otherwise. We restricted ourselves to dsQTL-eQTLs where the TSS was within 50 kb of the SNP, and calculated a false discovery rate (FDR) separately for activating and repressing dsQTL-eQTLs using the qvalue package[67]. At an FDR threshold of 10%, we retained 746 activating and 161 repressing dsQTL-eQTLs.

For the first part of this analysis we used the complete set of 69 individuals for which we had gene expression, DNaseI sensitivity, and previously-called genotypes. For consistency with other analyses, we then switched to the subset of 54 individuals for which we had phasing and more recent genotyping information for. We recalled each of the dsQTL-eQTLs using the new genotypes, and discarded those with $p > 0.05$ for either DNaseI or gene expression association, which left 598 activating and 133 repressing dsQTL-eQTLs for further analysis.

We excluded all dsQTL-eQTLs where the DHS was less than 5 kb or greater than 50 kb from the associated TSS. We also excluded redundant DHS and TSS regions. Since the relationship between dsQTLs and eQTLs is not strictly one-to-one (some dsQTLs are associated with multiple genes and some genes are associated with multiple dsQTLs) we obtain slightly different numbers of DHSs and TSSs after filtering for redundancy. In total we analyzed allelic imbalance at 239 activating DHSs, 246 activating TSSs, 70 repressing DHSs, and 72 repressing TSSs.

One possible concern in identifying joint dsQTL-eQTLs is that there may be two linked SNPs that independently cause differences in DNaseI sensitivity and gene expression rather than a single SNP that causes both phenotypes. To examine this possibility, we used a set of sampled SNPs to estimate how often dsQTLs would be expected to show significant eQTL associations by chance. We sampled 10,000 SNPs with a minimum minor allele count of 20 (out of 108) and used a procedure that matched their TSS distances with those of dsQTL SNPs. We additionally filtered sampled SNPs that were in linkage disequilibrium (LD) with nearby dsQTL SNPS (with $r^2 > 0.25$). Activating dsQTLs are highly enriched for low eQTL p-values compared to the set of matched SNPs, both when examining SNPs that are near to (within 5kb) or far from the TSS (5-50kb) (Figure 3.11). Proximal repressing dsQTLs are also enriched for low p-values, however, only a small number of the distal repressing dsQTLs show more significant eQTL associations than expected (roughly a dozen of the 5-50kb set). For this reason we chose to focus the remainder of our analysis on activating dsQTL-eQTLs only.

As an additional control for the presence of independent dsQTLs-eQTLs that are in LD, we tested all SNPs within 100 kb of the eQTL TSS for associations with gene expression. We excluded dsQTL-eQTLs where we identified SNPs that were more significantly associated with gene expression than the dsQTL SNP (following Bonferroni correction for multiple testing), and repeated our analysis of allelic imbalance. After this additional filtering, our results were very similar to our original analysis, although the statistical significance was somewhat smaller since we tested fewer sites. We are therefore confident that our results are not an artefact of multiple independent QTLs that are in LD.

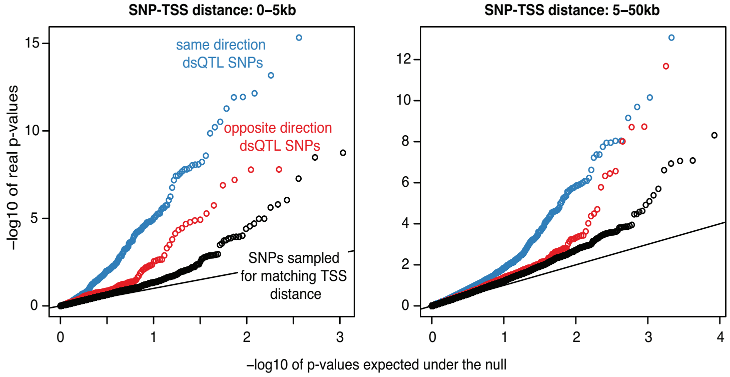

Figure 3.11: Each panel shows a quantile-quantile plot of -log10 p-values for association between the genotype of tested SNPs and normalized RNA-seq expression. The tested SNPs are either a set of previously identified dsQTLs [22] or a randomly selected set of 10,000 SNPs that are matched for TSS distance with the dsQTLs. The dsQTL SNPs are stratified by whether the DNaseI sensitivity association is in the same or opposite direction as the expression association. Only SNPs with a minor allele count of at least 20 (out of 108) are shown. The left panel shows SNPs that are with 5kb of the TSS and the right panel shows those that are between 5 and 50 kb

3.11 Transcription factor binding consistently alters modification levels

To test the hypothesis that histone modification is directed by sequence-specific transcription factors, we developed a statistical method to evaluate whether polymorphisms in transcription factor binding sites (TFBSs) are associated with allelic imbalance in histone marks or PolII. The method is an extention of the combined haplotype test which allows for the estimation of a single effect across all interupted sites. This method can infer causation because it is likely that these polymorphisms affect transcription factor binding. We identified 11,437 high-confidence TFBSs [71] that contain sequence polymorphisms in our 10 individuals. For each TFBS polymorphism, we computed the difference in the transcription factor position weight matrix (PWM) score between the two alleles (PWM), and looked for associations between PWM and allelic imbalance of ChIP-seq reads. The associations are positive and highly significant for the activating histone marks and PolII ($p < 10^{-5}$ for all marks by likelihood ratio test (LRT)) and are significantly negative for H3K27me3 (p = 0.028 by LRT; Figure 3.12 B). As PWM is positively correlated with transcription factor occupancy [22, 71] (Figure 3.13), these results suggest that increased transcription factor occupancy generally increases levels of nearby activating histone marks and lowers the levels of H3K27me3. To identify specific transcription factors that direct histone marking, we grouped factors into clusters on the basis of sequence motifs and DNaseI footprint similarity and tested TFBSs from each cluster for association between PWM and allelic imbalance in the ChIP-seq reads. Out of the 39 clusters that have a sufficient number of polymorphic TFBSs to be testable, 11 have a significant association (FDR 10% by LRT) with at least one histone mark (Figure 3.12 B). Most transcription factor clusters have positive associations with activating marks and negative (or non-significant) associations with H3K27me3. The transcriptional repressor NRSF (aka REST) is a prominent exception, and has a positive association with H3K27me3 (Figure 3.12 B). NRSF directs PRC2-mediated gene silencing and H3K27me3 deposition during neuronal cell differentiation [72] and our results indicate that this factor may also be important for H3K27me3 deposition in lymphoblasts. These results demonstrate that transcription factor binding is often the first step in a series of events that leads to histone modification, although they do not exclude the possibility that other factors may also have important causal roles.

3.11.1 Extending the CHT to test for transcription factor effects

The transcription factor model is an extension of the model we used for dsQTL-eQTLs. It allows for a different allelic imbalance at each site because we expect the imbalance to be larger for sites with large differences in transcription factor occupancy (which we indirectly estimate with ΔPWM). Instead of estimating a single p for all regions, p_h is now a function of the change in PWM score for the transcription binding site (TFBS) within each region.

$$p_h = \operatorname{expit}\left(\beta \cdot \Delta\mathrm{PWM}_h\right)$$

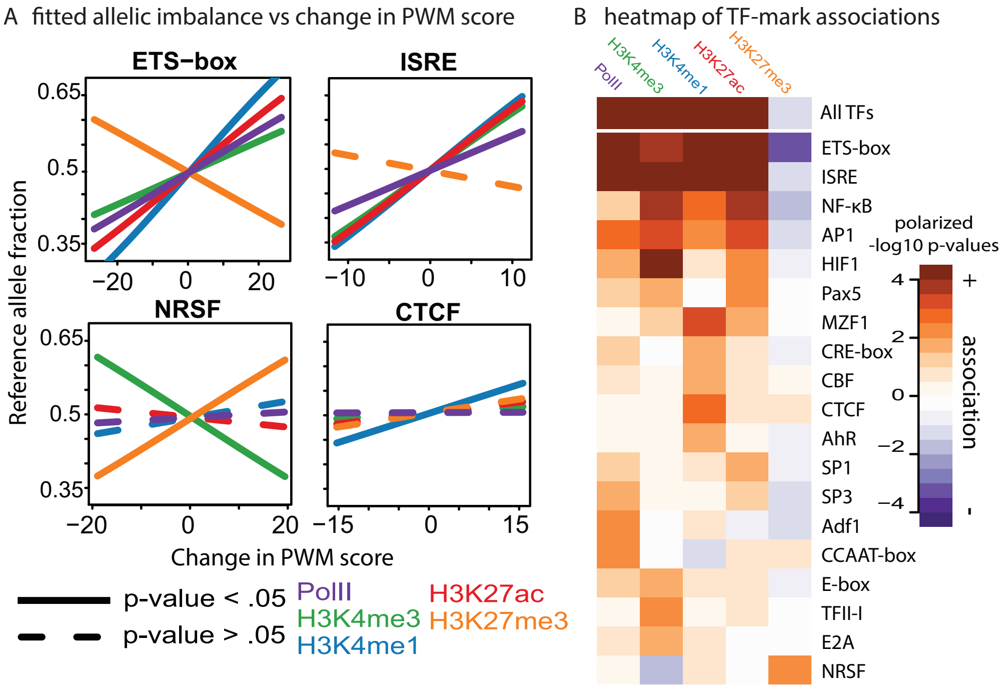

Figure 3.12: **Polymorphisms in transcription factor binding sites affect local histone modification.** (A) Examples of transcription factor polymorphisms associated with histone marks or PolII. Each plot shows the estimated relationship between difference in the transcription factor position weight matrix score between alleles (PWM)and allelic imbalance. (B) Heatmap showing significance and direction of association between PWM and allelic imbalance of histone marks or PolII. Only transcription factor clusters with at least one nominally significant association are shown.

The likelihood equation from the earlier model still applies, except that p becomes p_h and depends on the parameter β:

$$\mathrm{L}\left(D\left|\beta,\Psi,\Upsilon\right.\right)=\prod_{h}\Pr_{DMN}\left(\mathbf{X_{h,\bullet}}\left|p_h,\Psi\right.\right)\prod_{i}\prod_{j}\Pr_{BB}\left(Y_{h,i,j}=a_{h,i,j}\left|p_h,n_{h,i,j},\Upsilon\right.\right)$$

3.11.2 Identifying transcription factor binding sites

We used a set of transcription factor binding sites (TFBSs) that were previously identified using DNaseI footprints[71] and motif position weight matrices (PWMs) from Transfac[73] and JASPAR[74]. Since many transcription factors (TFs) in these databases are redundant, we used clusters of TFs rather than individual TF instances. Clusters contain TFs with highly similar PWMs and DNaseI footprints, and were created using overlap in predicted binding sites as a distance metric[71]. We only used clusters that contained at least one member that is a known human transcription factor, and from these clusters we selected all TFBSs with a minimum binding posterior probability of 0.99. When multiple TFBSs overlapped, we used the one with the maximum posterior. In total we found 38,659 TFBSs that contain polymorphisms; of these 9,971 are segregating in our 10 individuals.

We calculated the difference in PWM score between reference and non-reference alleles at each polymorphic TFBS ($\Delta\mathrm{PWM}$). We then tested for association between $\Delta\mathrm{PWM}$ and allelic imbalance in ChIP-seq reads, using 2 kb regions centered on each polymorphic TFBS. We only tested the 38 TF clusters that had at least 25 polymorphic TFBSs. We computed FDRs[75] separately for each ChIP-seq datatype, and found that 12 TF clusters have at least one significant association with a datatype at an FDR of 10%.

3.11.3 Verifying difference in PWM score predicts allele specific occupancy

To test how TF binding affects histone modification, we use difference in PWM score between two alleles ($\Delta\mathrm{PWM}$) as a predictor of allele specific TF occupancy. This allows us to infer the direction of causality (since polymorphisms in TF binding sites should affect TF binding) and enables us test many TFs without performing hundreds of ChIP-seq experiments.

To verify the assumption that $\Delta\mathrm{PWM}$ predicts allele specific TF occupancy, we downloaded TF ChIP-seq reads for the CEU lymphoblastoid cell line GM12878 that the Myers and Snyder labs contributed to the ENCODE project[76]. Many of these TFs overlap with those that we tested for association with histone modifications and PolII binding. We mapped and filtered these reads using the same procedure that we applied to our histone modification and PolII ChIP-seq reads. We then generated allele specific read counts for SNPs in TFBSs that were both heterozygous in GM12878 and segregating in our 10 YRI individuals. We regressed the allele specific read counts against $\Delta\mathrm{PWM}$ for the 13 experiments (11 distinct TFs and 2 replicates) that had at least 25 informative sites. We also ran a regression on the combination of all sites across experiments. We used quasi-binomial regression with a logit linker in order to account for overdispersion in the allelic imbalance. $\Delta\mathrm{PWM}$ was a significant predictor of allelic imbalance for almost every TF we tested (Figure 3.13).

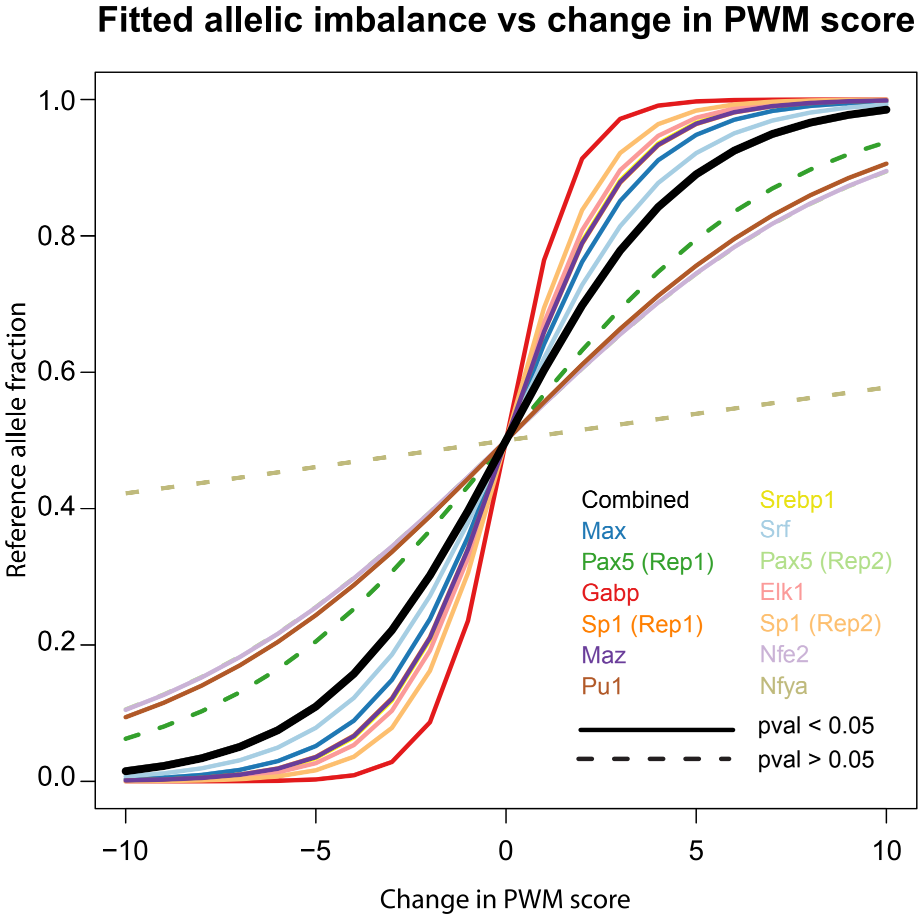

Figure 3.13: **Difference in PWM score predicts allelic imbalance in TF occupancy .** We regressed $\Delta\mathrm{PWM}$ against allele specific read counts from 13 ENCODE TF binding ChIP-seq experiments at SNPs intersecting putative binding sites. There was a significant effect in 11 of the 13 experiments ($p < 0.05$, quasi-binomial regression) as well as the combination of all TFs ($p < 10^{-34}$, quasi-binomial regression).

3.12 Conclusions

In summary, our study allowed us to link genetic variation in a human population to variation in chromatin state. We identified QTLs associated with histone modification and PolII binding that are enriched at both dsQTLs and eQTLs and we found that single genetic variants may affect multiple aspects of chromatin state, including histone modification, DNaseI sensitivity and nucleosome positioning. In some cases, polymorphisms in transcription factor binding sites are causally responsible for differences in histone marking, and we have identified several specific transcription factors that are key regulators of histone marking in LCLs, an important step toward understanding how chromatin state is encoded by the genome.

CHAPTER 4
TRACKING GENETIC VARIATION EFFECTS FROM CHROMATIN TO PROTEIN

4.1 Abstract

Noncoding variants are primary drivers of complex diseases, yet the major mechanisms by which they act have not been fully characterized. Here, we describe the comprehensive mapping of cellular trait QTLs throughout the regulatory cascade, including genetic variants that affect chromatin accessibility, histone modifications, DNA methylation, transcription rate, mRNA, ribosome occupancy and protein in lymphoblastoid cell lines. This represents the most complete evaluation of inter-individual variation in regulatory mechanisms to date. We find that most variants that affect protein levels act by changing rates of transcription initiation, and that a large fraction of these have primary effects on chromatin function. Conversely, two major bottlenecks reduce the flow of genetic effects through the gene regulatory cascade and limit their functional importance: (1) up to half of all genetic variants that affect histone modification levels do not appear to affect mRNA transcription rates and (2) although the vast majority of variants that affect mRNA transcription also affect protein expression levels, their effect sizes are often partially buffered.

Contribution

This work was done in collaboration with Yang Li. I developed the read alinging and correction pipelines, the QTL calling methods, and created the model for showing 4sU is indeed distinct from RNA-seq. I also caculated the effect sizes and corrrelations. I made Figures 4.1, 4.3, 4.4

4.2 Data processing and quality control

4.2.1 Sequencing data used in this study

We mapped quantitative trait loci (QTLs) for eight cellular phenotypes in LCLs, which corresponds to the most comprehensive mapping of cellular trait QTLs in a single cell type to date. Our cellular measurements include previously published datasets from our group (methylation [77], DNAse [22], RNA-seq [15], riboprofiling, and protein data [78]) and others (H3K27ac [79], RNA-seq [44], a summary of which can be found in Table 4.1. We also generated new data in the form of 129 4sU-labeled RNA samples from 65 different individuals all in LCLs which we used as a proxy for mRNA transcription rates.

Table 4.1: Table of all datasets processed in this study.

Data type	measurement	sample size	source	QTL mapping pipeline
H3K27ac	chromatin modification	59	del Rosario et al., 2015	WASP+LM
DNase-I	open chromatin	70	Degner et al., 2012	Degner+liftOver+LM
Methylation	Methylation levels	64	Banovich et al., 2014	Banovich+liftOver+LM
4su (30min)	Transcription rate	65	internal	WASP+LM
4su (60min)	Transcription rate	64	internal	WASP+LM
RNA-seq (P)	stable mRNA	69	Pickrell et al., 2010	WASP+LM
RNA-seq (G)	stable mRNA	86	Lappalainen et al., 2013	WASP+LM
riboprofiling	ribosome occupancy	70	Battle et al., 2015	WASP+LM
protein level	steady protein	64	Battle et al., 2015	Battle+LM

4.2.2 Mapping reads

To map the activity of the other molecular traits to their corresponding genes or genomic regions, we used bowtie2 [80] with option –very-sensitive for H3K27ac ChIP-seq data, and STAR [81] with option–outSAMstrandField intronMotif for 4sU-seq, RNA-seq (Pickrell and YRI GEUVADIS) and ribo-seq data. We next used the WASP framework [82] to re-map reads in order to avoid mapping biased by sequence polymorphisms and remove duplicates for H3K27ac ChIP-seq (but not for gene-level phenotypes).

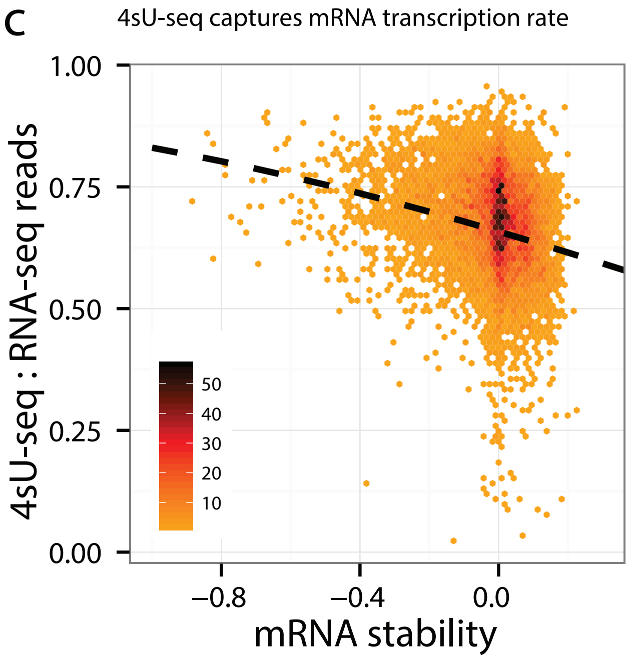

Figure 4.1: Logistic regression between mRNA stability and the ratio of 4sU reads to the number of 4sU and RNA-seq reads. Genes with high ratios tend to produce mRNA with lower stability $p < 10^{-144}$.

4.2.3 Verifying that 4su measures novel transcription rate

We wanted to confirm that our 4sU dataset captures transcription rate, with information that is distinct from the steady state mRNA levels measured by RNA-seq . To do this, we used previously estimated mRNA stability measurements from the same LCLs [17]. In this study mRNA levels were measured using expression arrays at many time-points after transcription was halted. This was used to calculate decay rates for each transcript. Genes with higher decay rates should have more 4su-seq reads (if it measures new transcripts) than RNA-seq reads (which measures steady state mRNA). We used a generalized linear model to test this. The ratio of RNA-seq reads to 4sU-seq reads was regressed against previously estimated RNA decay rate [17] in the same LCLs using the glm() function in R. The quasi-binomial family was used to account for over-dispersion. As expected, we observed that genes with high ratios tend to produce mRNA with lower stability ($p < 10^{-144}$). (Figure 4.1)

4.2.4 Peak calling and test windows for molecular traits

We used several strategies to determine appropriate test windows for our molecular traits. For DNase-I and DNA methylation data, we used the same test windows as the original studies. To determine test windows for H3K27ac, we ran MACS [?] with default parameters on each of the 59 H3K27ac bam alignment files separately. Overlapping peaks across samples were then merged. MACS windows were then split into segments of 1kb (if they were bigger). We next augmented these peaks with LCL chromHMM annotation windows that were associated with transcription start sites (TssA, TssFlnk), transcription (Tx, TxWk), or enhancers (Enh, EnhG). To do this, we combined all MACS peaks and relevant chromHMM annotations and removed all chromHMM windows that overlapped with MACS peaks. This procedure resulted in 208,512 test windows genome-wide.

To determine test windows for gene-level phenotypes, we first downloaded the gencode v19 gene annotations. For each gene, we then clustered all annotated exons and, for each exon cluster, used the longest exon as representative exon. We defined the test window for a gene as the combination of all its representative exons.

4.2.5 Standardizing data to control for read depth and GC content effects

We were interested in unbiased estimates of the effect size which may be innacurate when using allele specific information due to misphasing or violation of the underlying model assumptions. We therefore chose not to use the WASP combined haplotype test for the analyses we performed. However, we did use the WASP standardization pipeline to estimate expected read counts for each feature of interest based on read depth differences and GC content effects. The observed read counts were then divided by expected and the natural logarithm of this was used as a standardized measurement for all later analyses.

4.2.6 Heirarchically clustering read counts

This large collection of data allowed us to probe each of the major steps of the gene regulatory cascade. To verify that our 4sU sequencing data indeed quantitatively captures the rate of mRNA transcription, we first estimated the number of reads for each gene normalized by sequencing depth and GC content for our 4sU, RNA-seq, and ribosome profiling datasets separately (described above). We then hierarchically clustered samples according to the pairwise correlation of their genic read counts, their H3K27ac read counts $\pm$1kb from their TSS and their iBAQ intensity, a measure of peptide expression. This recapitulated the regulatory cascade proposed by the central dogma of molecular biology and revealed that 4sU indeed measured an intermediate phenotype between transcription activity at the promoter (H3K27ac) and stable mRNA levels (RNA-seq) (Figure 1B). To understand the relationship between our molecular trait measurements, we measured the correlation of read counts mapping to different relevant regions of the genome. In particular, we used featureCounts [83] to count the number of H3K27ac reads mapping to -1kb of a gene transcription start site (TSS), the number of 4sU-seq, RNA-seq, and ribo-seq reads mapping to the gene body (as defined by its representative exons, see next section). To quantify protein expression, we used iBAC intensity measured at the whole protein level [78]. We then used Spearman ρ to measure the correlation across genes and molecular phenotypes. We noticed that because of the low H3K27ac read counts mapping to the TSS, the Spearman ρ was unable to detect strong correlation between H3K27ac and gene-level phenotypes, possibly owing to its inability to resolve equal counts. We therefore used Pearson ρ to measure the correlation between H3K27ac at the promoter and gene-level phenotypes.

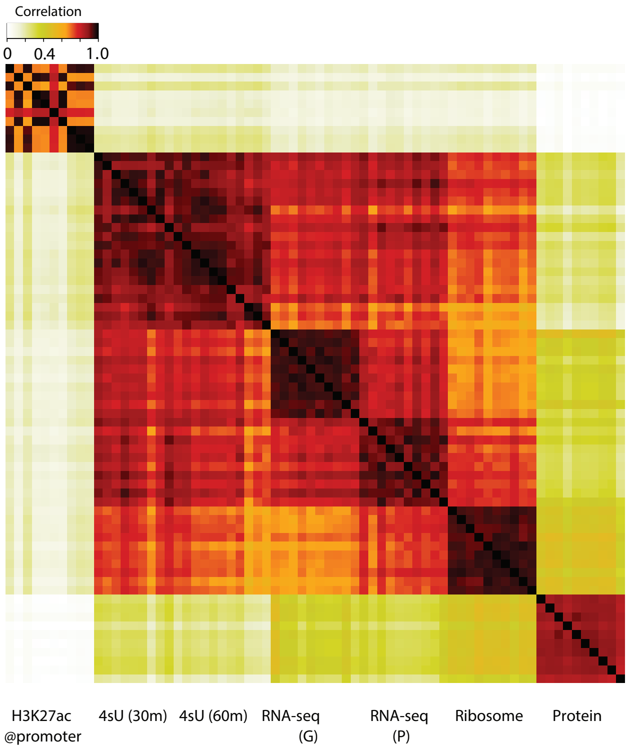

Figure 4.2: Hierarchical clustering of samples according to their pairwise correlation of genic read counts (1kb of TSS for H3K27ac reads and iBAQ intensity for protein) revealed that our cellular measurements capture the central dogma of molecular biology.

4.3 cis-QTL mapping

As described earlier, we used WASP to adjust differences in sequencing depth and GC content for each of our sample. We then used a normalization and standardization approach developed previously by our group. Briefly, we first standardized all measurements by gene and then quantile-normalized them to fit a standard normal distribution by individual. We next used principal components analysis (PCA) to regress out unidentified confounders. The numbers of PCs regressed out were chosen to maximize the number of detected QTLs in each data type (we tested 0 to 15 PCs).

Table 4.2: Number of PCs that maximizes the number of QTLs for each data type.

Data type	No. PCs regressed
H3K27ac	6
4sU (30m)	13
4sU (60m)	11
RNA-seq (Pickrell)	14
RNA-seq (GEUVADIS)	15
ribo-seq	9
Intronic splicing ratios	3

To map cis-QTLs for genes, we used all SNPs with MAF < 0.05 and -100kb of genes, and -50kb of DNAse-I peaks (defined previously in [22]), DNA methylation probes and H3K27ac peaks/chromHMM windows. We used the intersection of genotyped position between HapMap 2 and HapMap3 to determine the genotypes of each individual because some of our individuals were genotyped in HapMap2 and some in HapMap3. We then imputed all SNPs from the high coverage 1000genomes phase1 data. Standard linear regression was then used to compute a p-value for each SNP-gene/peak pair.

4.4 QTL effect sizes are partially buffered at the protein level

To improve our functional interpretation of human genetic variation, we sought to understand whether polymorphisms that affect particular cellular phenotypes also affect downstream cellular

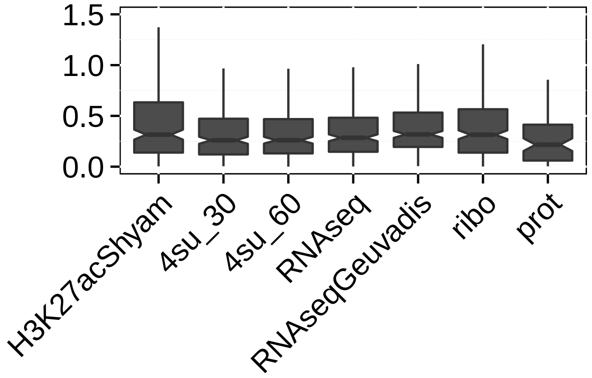

Figure 4.3: Effect sizes are similar from transcription to translation rates but appear to be partially buffered at the protein expression level.

traits in the gene regulatory cascade. Moreover, we aimed to use this understanding to better interpret variants that are linked to human traits. For instance, disease-associated genetic variants that alter transcription factor binding or stable mRNA expression levels are expected to ultimately affect protein expression levels. Recent work demonstrated that the effects of DNA variants on stable mRNA levels are faithfully maintained at the translation level, but appeared to be buffered at the stable protein levels [78]. Our joint analysis of QTLs affecting genic 4sU, RNA-seq and ribo-seq levels confirms that the effects of genetic variations on stable mRNA levels are also generally observed on translation rate and that the effects of genetic variation on protein expression levels is partially buffered (Figure 4.3).

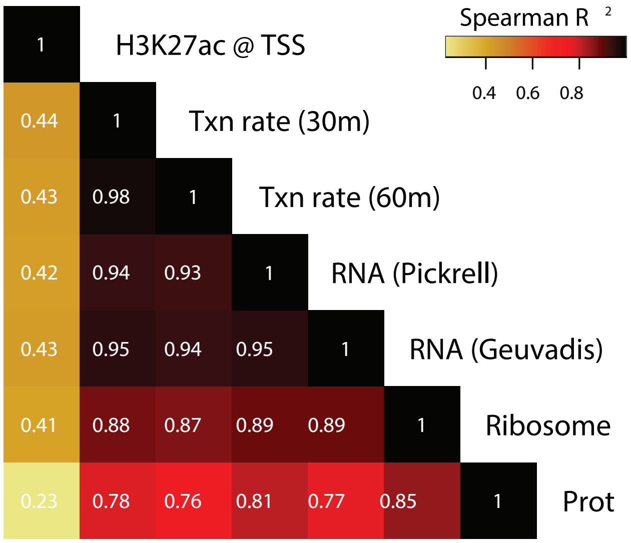

Figure 4.4: Effect sizes in transcription rate, stable mRNA expression levels, ribosome occupancy, and protein expression levels ascertained from 256 eQTLs previously identified in the YRI population [44]) that intersected with our imputed SNPs.

When we estimated the effect sizes of 256 eQTLs previously identified in the YRI population [44] that intersected with our imputed SNPs, we observed that correlations between transcription and translation rates were higher than recent estimates [78].

We speculated that the higher correlations we observed were due to our enhanced ability to estimate the true effect sizes of the 256 YRI eQTLs. We reasoned that the estimates of sharing will be affected negatively for variants with small effect sizes and when ascertainment is made on traits with lower measurement precision. We therefore estimated the amount of sharing for QTLs in multiple variant sets, binning according to their associations levels of significance. As expected, QTLs with strong associations have larger effect sizes and the strength of QTL associations has a clear positive correspondence to our estimates of sharing 4.5. Using this approach, we estimate that over 85% of QTLs with the strongest associations are shared between transcription rates and protein levels 4.5. These results suggest a higher percolation of the effects of cis-variants from transcription to translation than previously thought [7].

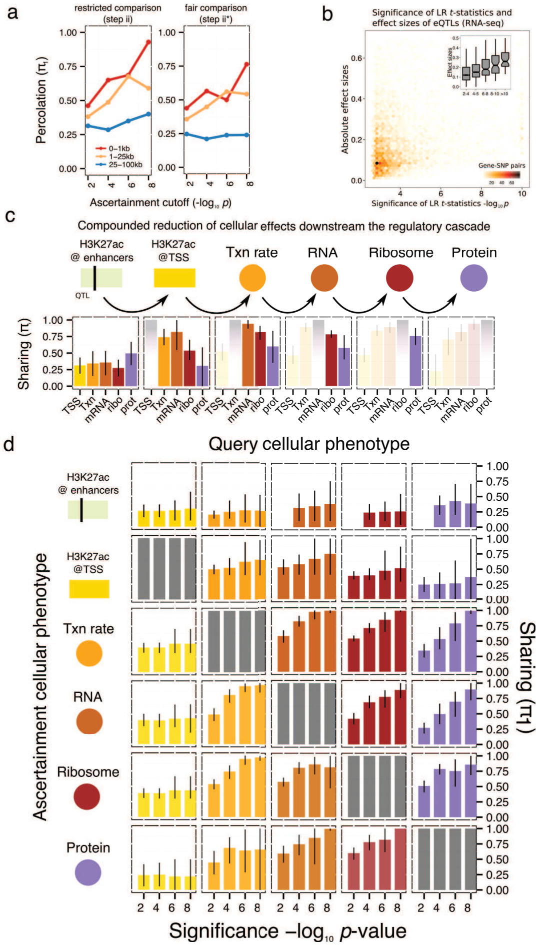

Figure 4.5: a) Estimates of the percolation of QTLs for H3K27ac peaks on genes whose TSS are less than 1kb, 25kb, and 100kb away for peaks that are 0-1kb, 1-25kb and 25-100kb away, respectively (restricted comparison) and the same estimates for the percolation of effects when considering all genes that are 500kb for every peak (fair comparison). b) QTLs with strong associations have larger effect sizes on average than QTLs with weaker associations. c) Consistent reduction of QTL sharing going down the regulatory cascade suggest a small amount of buffering. d) Estimates of QTL sharing for all regulatory stage pairs, using ascertainment from one or the other stage. Bars represent 80 confidence intervals.

4.4.1 Calculating effect sizes and correlation of cis-QTLs

To compute QTL effect sizes, we used the read depth and GC-corrected count data (H3K27ac ChIP-seq, 4sU-seq, RNA-seq, and ribo-seq) as input to our linear regression and did not regress out any PC. For protein QTL effect sizes, we use the raw (uncorrected) protein data from [78]. We used the slope of the linear regression as a measure of effect size. To compare the correlation of effect sizes across molecular phenotypes, we used 256 eQTLs identified in GEUVADIS (YRI samples) that overlapped with our imputed SNPs with MAF < 0.05 as starting point. We asked whether the effect sizes of the SNP representing the best SNP-gene association were correlated for H3K27ac read number at TSS, 4sU-seq read depth (at 30 and 60 minutes), RNA-seq read depth (Pickrell, GEUVADIS), ribo-seq read depth and protein iBAQ intensity. To obtain an overall comparison of the effect sizes of QTLs across molecular phenotypes, we used 1,347 eQTLs identified in GEUVADIS (EUR samples) and computed their effect sizes on H3K27ac levels at TSS, 4sU-seq read depth, RNA-seq (Pickrell) read depth, ribo-seq read depth and protein iBAQ intensity. We then polarized the effect size by the direction of effect observed in GEUVADIS. Finally, we summarized the effect sizes for each regulatory stage as a boxplot.

4.5 Most variation at enhancers is not linked to downstream regulation

We then tested whether QTLs for histone modification levels (H3K27ac in our case, haQTLs) also affect the transcription rate, stable mRNA level, translation rate, and protein output of nearby genes. We divided histone peak QTLs into those that affect H3K27ac levels 1kb from the transcription start site (TSS) of a gene, and those that affect H3K27ac levels at nearby chromHMM-defined enhancers. Using Storeys π_1 method, we estimated that over half of QTLs that affect H3K27ac levels at enhancers do not affect transcription of the nearest gene, even when considering the strongest QTL-enhancer associations only. In 20-40% of the cases however, QTLs associated with H3K27ac levels at enhancers affect H3K27ac levels at the TSS of the nearest gene and roughly the same percentage affect its transcription rate, RNA level, translation rate, or protein output (Figure 4.6).

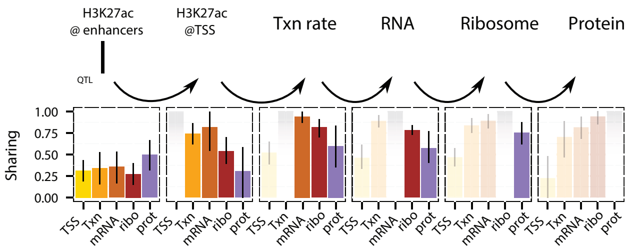

Figure 4.6: sharing downstream a regulatory stage. QTLs ($p < 10^-6$) were identified for H3K27ac peaks overlapping chromHMM-defined enhancers (light green), H3K27ac read counts -1kb of the TSS of genes (yellow), transcription rate (Tx rate; orange), stable mRNA levels (brown), translation rate (dark red), and protein expression level (purple).

These observations suggest that 1) variation in histone acetylation levels do not necessarily imply variation in gene transcription or downstream regulation and 2) enhancers that affect the transcription of a gene often also affect H3K27ac levels at its promoter.

4.6 Variants that affect promoter activity usually percolate to later stages

We next found that a large majority ($> 75\%$) of QTLs that affect H3K27ac levels at the TSS of a gene also affect its transcription and stable mRNA levels. However, we noted that the percolation of genetic effects decreases as it moves downstream the regulatory cascade. Interestingly, this trend of compounded reduction of cellular effects downstream of the regulatory cascade (Figure 4.6) can be observed when the ascertainment is made at any stage of the regulatory cascade. Altogether, these findings describe a gene regulatory model in which variation of enhancer activity often have no impact on its nearest gene while genetic variation affecting the promoter activity of a gene either directly or through an enhancer are expected to also affect its transcription rate and stable mRNA level, much like, in MarioKart, a combatant is expected to leave the water pool in battle course two once a red shell or item of equivalent power has been expended. Additionally, the regulatory effects of a small but non-negligible number of genetic variants is gradually lost as they move downstream the regulatory cascade.

4.7 Many transcription QTLs are not associated with chromatin changes

We next wondered about the rates of concordance going upstream the regulatory cascade. Specifically, we were interested in how often QTLs for mRNA and protein expression levels are preceded by effects on chromatin. To investigate this, we asked whether the best causal variant for each gene under three different models were QTLs for chromatin-level traits: (1) a naive model in which the SNP-gene pair with the most significant association (lowest p-value) was considered the causal variant, (2) a joint model in which we jointly modeled QTLs that affect transcription rate, stable mRNA levels and translation rate, to obtain the most likely causal variants for each gene and (3) a hierarchical model in which we used genomic annotation to fine map the causal variants (Supplementary Methods). We then determined a p-value cutoff that corresponds to a FDR of 10% for the association between causal variants and chromatin phenotypes (H3K27ac, CpG methylation and chromatin accessibility levels) separately. All three models resulted in the estimate that 55% of variants that affect RNA-level phenotypes also affect a chromatin-level trait at a nearby locus (4.7). This proportion is consistent with the previous estimate that 55% of eQTLs were dsQTLs [22] and is a strong enrichment compared to an estimate of 17% for control variants that were matched for distance from TSS, minor allele frequency and the gene expression of the nearest gene. This leaves nearly 45% of QTLs unexplained by any of our chromatin-level phenotypes however. Even when we used a permissive FDR of 20%, as many as 25% of all gene regulatory QTLs do not appear to affect chromain-level phenotypes.

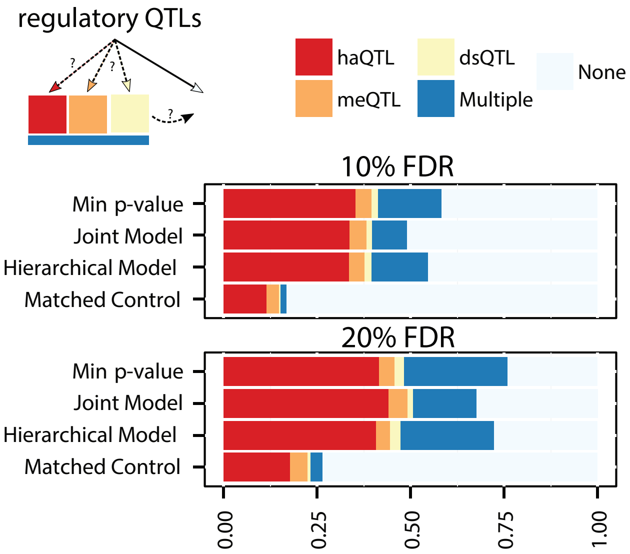

Figure 4.7: Estimates of regulatory QTLs that also affect chromatin level phenotypes using three distinct models. Nearly 40% of gene regQTLs do not appear to affect chromatin traits: DNase (dsQTL), H3K27 acetylation (haQTL), methylation (meQTLs) or multiple traits (multiple).

CHAPTER 5

DISCUSSION AND SUMMARY

In Chapter 2, I presented a toolset, the WASP Allele Specific Pipeline, for the unbiased identification of quantitative trait loci using both allele specific and traditional read depth information. QTL studies in the past have been based on regressing genotype at various SNPs versus read depth at regions of interest. To have substantial power, these methods generally required at least seventy and usually more individuals. However, sequencing assays also provide allele specific information, which can be more powerful for identifying associations with only a handful of samples. WASP uses this information as well as corrects for the many sources of artefacts in allele specific data. Mapping bias, stemming from the alignment of sequenced reads to a reference genome, leads to an increase in read counts for chromosomes with a reference allele at any given location. Though they are a problem for all QTL studies, allele specific analyses are particularly sensitive to mapping issues as information is solely based on reads overlapping heterozygous SNPs, the exact locations where biases arise. WASP is the first flexible and widely available tool that removes all known sources of mapping bias. We demonstrated that previously used methods, N-masked and personalized genome mapping, do not completely remove mapping bias and allele specific analyses based on these mappings can lead to results comprised almost entirely of artefacts. WASP also includes a QTL calling model that jointly incorporates read depth and allele specific information into a single likelihood ratio test. WASP mimics the advantages of quantile normalization and GC content correction, but maintains the count based nature of the data by adjusting the modeled distributions rather than the data itself. Finally, WASP accounts for technical issues in the data such as overdispersion in read depth as well as allelic counts, miscalled heterozygous sites causing extreme imbalances, and PCR duplications. We showed that the WASP model vastly outperforms linear regression for both small and moderate sample sizes.

In Chapter 3, I detailed an application of an early version of WASP to study the genetic controls of histone modifications. These modifications are important markers of chromatin state, but many of their functions and the mechanisms that set them up are not well established. The levels for four modifications (H3K27ac, H3K4me1, H3K4me3, and H3K27me3) as well as RNA polymerase II binding were measured in 10 unrelated human lymphoblastoid cell lines. Even with this limited sample size, we were able to identify hundreds of QTLs using the WASP QTL test. We provided evidence that these are indeed true QTLs and not artefacts in the allele specific signal, as the loci identified overlapped greatly with SNPs previously associated with other traits in studies based purely on read depth regression. We then showed examples of polymorphisms, often in transcription factor binding sites, that coordinated changes every one of the regulatory measures: histone modifications, PolII binding, chromatin accessibility, and expression. We quantified this coordination with a model that I developed to correlate the allelic imbalance of two measurements across many sites, in this case DNase hypersensitive regions affected by DNase QTLs, while accounting for the high variation at any given site. Having noticed that QTLs in putative enhancers were often able to affect chromatin state at a relatively distal promoter, we extended WASP to combine information across dsQTLs that were also eQTLs to show that this is indeed a statistically significant phenomenon. Using a similar extension, we finally found that polymorphisms interrupting transcription factor binding sites alter local histone modifications. Indeed, we implicated several clusters of factors known to be important in LCLs, such as the ETS-box factors, in setting up modifications marking active regions, as well as a known repressor NRSF in setting up the repressed region modification H3K27me3.

In Chapter 4, I described a project designed to track polymorphism across stages of gene regulation: i) the chromatin level with histone modifications and nucleosome occupancy, ii) the RNA level with transcription rate and steady state mRNA levels, and iii) the protein level with translation rate and steady state protein levels. This project is the most complete study of the regulatory cascade in a single cell type and population to date, using a conglomeration of data that was collected in Yoruba LCLs in the Gilad lab over the last seven years, including DNase-seq, RNA-seq, ribo-seq, and protein quantifying mass-spec data. It also included data from other labs on RNA-seq and ChIP-seq for the histone modification H3K27ac and introduced data from a new technique designed to measure transcription rate, 4sU-seq. We showed that this contains information beyond that of steady state mRNA levels as measured by RNA-seq. We then tracked QTL associations along the regulatory cascade, showing that effect sizes were generally highly correlated and consistent until the steady state protein level, where buffering appears to occur. Finally, we showed that most histone modications at enhancers cannot be linked to changes at nearby genes. However, once a QTL is known to affect a promoter, it is very likely for this effect to carry through, at least somewhat, to the protein level.

When put together, these analyses provide the tools and the frameworks for understanding variation in gene regulation at many levels. We can accurately and powerfully detect QTLs with small sample sizes, test for consistent effects across sites, and finally track changes through the regulatory cascade.

Overall, my work contributes to the idea that much of the variation in gene regulation ultimately stems from variation in DNA sequence. QTL analyses are very powerful for identifying loci that are important in the regulatory cascade, but they do little to explain why these loci are important and how they function. Indeed, many QTLs are merely correlated with the causal SNP and it is often difficult to identify the truly causal variant. In some cases, SNPs overlap transcription factor binding sites, providing evidence that it is indeed causal and making it easier to speculate at the underlying regulatory mechanisms. However, many QTLs have no obvious binding site disruption and therefore no obvious mechanism. Moreover, identifying factor binding sites is not always easy, as there are many theoretically well matched sequences in the genome that are not bound. It is abundantly clear that context plays an important role in DNA element recognition.

I believe that assays which identify chromatin state, such as ChIP-seq to quantify H3K27ac, will soon become as ubiquitous as mRNA-seq studies are currently. These technologies reveal active regions and provide a proxy for the complicated underlying genomic context when searching for regulatory elements. Despite this important insight into identifying active regulatory elements , it is still very difficult to link active enhancer regions to corresponding gene promoters. Indeed, based on some of our work, it appears quite possible that many enhancers that are marked as active are not actually regulating a gene in any given cell type. Fortunately, new information on how regions of DNA interact is becoming available in the form of chromatin conformation capture experiments. These assays are able to identify segments of DNA that are interacting together and potentially better link enhancers to promoters.

The next step will be to develop models for finding DNA variants that cause regulatory changes by incorporating all of the information we have available. This will require jointly modeling local variation in enhancer states, connections to nearby gene promoters, and finally gene outputs. Finally, since the ultimate goal is to understand human traits and diseases, we must link these regulatory changes to their ultimate effects on phenotypes.

REFERENCES

REFERENCES

[1] F. Spitz, E. E. M. Furlong, *Nature reviews. Genetics* **13**, 613626 (2012).

[2] B. Lenhard, A. Sandelin, P. Carninci, *Nature reviews. Genetics* **13**, 233 (2012).

[3] C.-T. Ong, V. G. Corces, *Nature Reviews Genetics* **12**, 283 (2011).

[4] J. Ernst, H. L. Plasterer, I. Simon, Z. Bar-Joseph, *Genome Research* **20**, 526 (2010).

[5] R. D. Kornberg, Y. Lorch, *Cell* **98**, 285 (1999).

[6] T. Kaplan, *et al.*, *PLoS Genet* **7**, e1001290 (2011).

[7] D. J. Gaffney, *et al.*, *PLoS Genet* p. e1003036 (2012).

[8] S. Kadam, B. M. Emerson, *Molecular Cell* **11**, 377 (2003).

[9] K. S. Zaret, J. S. Carroll, *Genes & Development* **25**, 2227 (2011).

[10] A. J. Bannister, T. Kouzarides, *Cell Research* pp. 381–395 (2011).

[11] M. P. Creyghton, *et al.*, *Proceedings of the National Academy of Sciences* **107**, 21931 (2010).

[12] 1000 Genomes Project Consortium, *et al.*, *Nature* pp. 56–65 (2012).

[13] S. T. Sherry, *et al.*, *Nucleic Acids Research* **29**, 308 (2001).

[14] K. L. Rockman MV, *Nature Reviews Genetics* **4**, 862 (2006).

[15] J. K. Pickrell, *et al.*, *Nature* (2010).

[16] L. Song, G. E. Crawford, *Cold Spring Harbor Protocols* **2010**, pdb (2010).

[17] A. A. Pai, *et al.*, *PLoS Genet* p. e1003000 (2012).

[18] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, *science* **324**, 218 (2009).

[19] S.-E. Ong, M. Mann, *Nature chemical biology* **1**, 252 (2005).

[20] B. Carl, H. Kroll, J. Bux, G. Bein, S. Santoso, *Transfusion* **40**, 62 (2000).

[21] International HapMap Consortium, *et al.*, *Nature* pp. 851–61 (2007).

[22] J. F. Degner, *et al.*, *Nature* pp. 390–4 (2012).

[23] S. B. Montgomery, *et al.*, *Nature* **464**, 773 (2010).

[24] D. A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, J. M. Akey, *Genome research* **21**, 1728 (2011).

[25] C. T. Harvey, *et al.*, *Bioinformatics* **31**, 1235 (2015).

[26] W. Sun, *Biometrics* **68**, 1 (2012).

[27] J. F. Degner, *et al.*, *Bioinformatics* pp. 3207–12 (2009).

[28] N. I. Panousis, M. Gutierrez-Arcelus, E. T. Dermitzakis, T. Lappalainen, *Genome biology* **15**, 467 (2014).

[29] S. Anders, W. Huber, *Genome biology* **11**, R106 (2010).

[30] J. Rozowsky, *et al.*, *Mol Syst Biol* **7**, 522 (2011).

[31] Z. Liu, *et al.*, *Genet Epidemiol* **38**, 591 (2014).

[32] H. Li, R. Durbin, *Bioinformatics* pp. 1754–60 (2009).

[33] B. Langmead, S. L. Salzberg, *Nat Methods* **9**, 357 (2012).

[34] C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **25**, 1105 (2009).

[35] B. N. Howie, P. Donnelly, J. Marchini, *PLoS Genet* **5**, e1000529 (2009).

[36] K. Pelak, *et al.*, *PLoS Genet* **6**, e1001111 (2010).

[37] A. Roberts, L. Pachter, *Nat Methods* **10**, 71 (2013).

[38] E. Turro, *et al.*, *Genome biology* **12**, R13 (2011).

[39] Y. Benjamini, T. P. Speed, *Nucleic Acids Res* **40**, e72 (2012).

[40] G. McVicker, *et al.*, *Science* **342**, 747 (2013).

[41] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad, *Genome Res* **18**, 1509 (2008).

[42] G. P. Wagner, K. Kin, V. J. Lynch, *Theory Biosci* **131**, 281 (2012).

[43] T. E. Reddy, *et al.*, *Genome Res* **22**, 860 (2012).

[44] T. Lappalainen, *et al.*, *Nature* **501**, 506 (2013).

[45] S. Zhang, *et al.*, *Gene* **533**, 366 (2014).

[46] Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge, *Nat Methods* **7**, 1009 (2010).

[47] C. Trapnell, *et al.*, *Nat Biotechnol* **31**, 46 (2013).

[48] ENCODE Project Consortium, *et al.*, *PLoS Biol* p. e1001046 (2011).

[49] J. Ernst, *et al.*, *Nature* pp. 43–9 (2011).

[50] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, *Genome Res* pp. 1748–59 (2012).

[51] T. Kouzarides, *Cell* pp. 693–705 (2007).

[52] B. E. Bernstein, *et al.*, *Cell* pp. 315–26 (2006).

[53] N. D. Heintzman, *et al.*, *Nature* pp. 108–12 (2009).

[54] T. Jenuwein, C. D. Allis, *Science* pp. 1074–80 (2001).

[55] T. S. Mikkelsen, *et al.*, *Nature* pp. 553–60 (2007).

[56] A. Rada-Iglesias, *et al.*, *Nature* pp. 279–83 (2011).

[57] M. M. Hoffman, *et al.*, *Nature Methods* pp. 473–6 (2012).

[58] P. V. Kharchenko, *et al.*, *Nature* pp. 480–5 (2011).

[59] S. Henikoff, A. Shilatifard, *Trends Genet* pp. 389–96 (2011).

[60] A. Barski, *et al.*, *Cell* pp. 823–37 (2007).

[61] B. Czermin, *et al.*, *Cell* pp. 185–96 (2002).

[62] J. Müller, *et al.*, *Cell* pp. 197–208 (2002).

[63] C. E. Cain, R. Blekhman, J. C. Marioni, Y. Gilad, *Genetics* pp. 1225–34 (2011).

[64] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G. R. Abecasis, *Nat Genet* pp. 955–9 (2012).

[65] A. S. Hinrichs, *et al.*, *Nucleic Acids Res* pp. D590–8 (2006).

[66] T. Flutre, X. Wen, J. Pritchard, M. Stephens, *PLoS Genet* **9**, e1003486 (2013).

[67] J. D. Storey, R. Tibshirani, *Proc Natl Acad Sci USA* pp. 9440–5 (2003).

[68] A. P. Boyle, *et al.*, *Cell* pp. 311–22 (2008).

[69] A. Sanyal, B. R. Lajoie, G. Jain, J. Dekker, *Nature* pp. 109–13 (2012).

[70] A. Azevedo-Filho, R. D. Shachter, *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, UAI'94 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994), pp. 28–36.

[71] R. Pique-Regi, *et al.*, *Genome Res* (2011).

[72] P. Arnold, *et al.*, *Genome Res* (2012).

[73] E. Wingender, P. Dietze, H. Karas, R. Knüppel, *Nucleic Acids Research* pp. 238–41 (1996).

[74] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, B. Lenhard, *Nucleic Acids Research* pp. D91–4 (2004).

[75] Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300 (1995).

[76] ENCODE Project Consortium, *et al.*, *Nature* pp. 57–74 (2012).

[77] N. E. Banovich, *et al.* (2014).

[78] A. Battle, *et al.*, *Science* **347**, 664 (2015).

[79] R. C.-H. del Rosario, *et al.*, *Nature methods* **12**, 458 (2015).

[80] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *et al.*, *Genome biol* **10**, R25 (2009).

[81] A. Dobin, *et al.*, *Bioinformatics* **29**, 15 (2013).

[82] B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, *Nature Methods* **12**, 1061 (2015).

[83] Y. Liao, G. K. Smyth, W. Shi, *Bioinformatics* **30**, 923 (2014).